Convex Relaxation for Combinatorial Penalties

Guillaume Obozinski, Francis Bach

Convex Relaxation for Combinatorial Penalties

Guillaume Obozinski
INRIA - Sierra project-team
Laboratoire d'Informatique
de l'Ecole Normale Supérieure
Paris, France
guillaume.obozinski@ens.fr

Francis Bach
INRIA - Sierra project-team
Laboratoire d'Informatique
de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

May 6, 2012

Abstract

In this paper, we propose an unifying view of several recently proposed structured sparsity-inducing norms. We consider the situation of a model simultaneously (a) penalized by a set-function defined on the support of the unknown parameter vector which represents prior knowledge on supports, and (b) regularized in ℓ_p-norm. We show that the natural combinatorial optimization problems obtained may be relaxed into convex optimization problems and introduce a notion, the *lower combinatorial envelope* of a set-function, that characterizes the tightness of our relaxations. We moreover establish links with norms based on latent representations including the latent group Lasso and *block-coding*, and with norms obtained from submodular functions.

1 Introduction

The last years have seen the emergence of the field of *structured sparsity*, which aims at identifying a model of small complexity given a priori knowledge on its possible structure.

Various regularizations, in particular convex, have been proposed that formalized the notion that prior information can be expressed through functions encoding the set of possible or encouraged supports[1] in the model. Several convex regularizers for structured sparsity arose as generalizations of the group Lasso (Yuan and Lin, 2006) to the case of overlapping groups (Jenatton et al., 2011a; Jacob et al., 2009; Mairal et al., 2011), in particular to tree-structured groups (Zhao et al., 2009; Kim and Xing, 2010; Jenatton et al., 2011b). Other formulations have been considered based on variational formulations (Micchelli et al., 2011), the perspective of multiple kernel learning (Bach et al., 2012), submodular functions (Bach, 2010) and norms defined as convex hulls (Obozinski et al., 2011; Chandrasekaran et al., 2010). Non convex approaches include He and Carin (2009); Baraniuk et al. (2010); Huang et al. (2011). We refer the reader to Huang et al. (2011) for a concise overview and discussion of the related literature and to Bach et al. (2012) for a more detailed tutorial presentation.

In this context, and given a model parametrized by a vector of coefficients $w \in \mathbb{R}^V$ with $V = \{1, \dots, d\}$, the main objective of this paper is to find an appropriate way to combine together *combinatorial penalties*, that control the structure of a model in terms of the sets of variables

[1] By support, we mean the set of indices of non-zero parameters.

allowed or favored to enter the function learned, with *continuous regularizers*—such as ℓ_p-norms, that control the magnitude of their coefficients, into a convex regularization that would control both.

Part of our motivation stems from previous work on regularizers that "convexify" combinatorial penalties. Bach (2010) proposes to consider the tightest convex relaxation of the restriction of a submodular penalty to a unit ℓ_∞-ball in the space of model parameters $w \in \mathbb{R}^d$. However, this relaxation scheme implicitly assumes that the coefficients are in a unit ℓ_∞-ball; then, the relaxation obtained induces clustering artifacts of the values of the learned vector. It would thus seem desirable to propose relaxation schemes that do not assume that coefficient are bounded but rather to control continuously their magnitude and to find alternatives to the ℓ_∞-norm. Finally the class of functions considered is restricted to submodular functions.

In this paper, we therefore consider combined penalties of the form mentioned above and propose first an appropriate convex relaxation in Section 2; the properties of general combinatorial functions preserved by the relaxation are captured by the notion of lower combinatorial envelope introduced in Section 2.2. Section 3 relates the convex regularization obtained to the latent group Lasso and to set-cover penalties, while Section 4 provides additional examples, such as the exclusive Lasso. We discuss in more details the case of submodular functions in Section 6 and propose for that case efficient algorithms and a theoretical analysis. Finally, we present some experiments in Section 7.

Yet another motivation is to follow loosely the principle of two-part or multiple-part codes from MDL theory (Rissanen, 1978). In particular if the model is parametrized by a vector of parameters w, it is possible to encode (an approximation of) w itself with a two-part code, by encoding first the support $\mathrm{Supp}(w)$ —or set of non-zero values— of w with a code length of $F(\mathrm{Supp}(w))$ and by encoding the actual values of w using a code based on a log prior distribution on the vector w that could motivate the choice of an ℓ_p-norm as a surrogate for the code length. This leads naturally to consider penalties of the form $\mu F(\mathrm{Supp}(w)) + \nu \|w\|_p^p$ and to find appropriate notions of relaxation.

Notations. When indexing vectors of $\mathbb{R}^d$ with a set A or B in *exponent*, x^A and $x^B \in \mathbb{R}^d$ refer to two a priori unrelated vectors; by contrast, when using A as an *index*, and given a vector $x \in \mathbb{R}^d$, x_A denotes the vector of $\mathbb{R}^d$ such that $[x_A]_i = x_i$, $i \in A$ and $[x_A]_i = 0$, $i \notin A$. If s is a vector in $\mathbb{R}^d$, we use the shorthand $s(A) := \sum_{i \in A} s_i$ and $|s|$ denotes the vector whose elements are the absolute values $|s_i|$ of the elements s_i in s. For $p \geq 1$, we define q through the relation $\frac{1}{p} + \frac{1}{q} = 1$. The ℓ_q-norm of a vector w will be noted $\|w\|_q = \big(\sum_i w_i^q\big)^{1/q}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we will denote by f^* is Fenchel-Legendre conjugate. We will write $\overline{\mathbb{R}}_+$ for $\mathbb{R}_+ \cup \{+\infty\}$.

2 Penalties and convex relaxations

Let $V = \{1, \ldots, d\}$ and $2^V = \{A \mid A \subset V\}$ its power-set. We will consider positive-valued set-functions of the form $F : 2^V \to \overline{\mathbb{R}}_+$ such that $F(\varnothing) = 0$ and $F(A) > 0$ for all $A \neq \varnothing$. We do not necessarily assume that F is non-decreasing, even if it would a priori be natural for a penalty function of the support. We however assume that the domain of F, defined as $\mathcal{D}_0 := \{A \mid F(A) < \infty\}$, covers V, i.e., satisfies $\cup_{A \in \mathcal{D}_0} A = V$ (if F is non-decreasing, this just implies that it should be finite on singletons). We will denote by $\iota_{x \in S}$ the indicator function of the set S, taking value 0 on the set and $+\infty$ outside. We will write $[\![k_1, k_2]\!]$ to denote the discrete interval $\{k_1, \ldots, k_2\}$.

With the motivations of the previous section, and denoting by $\mathrm{Supp}(w)$ the set of non-zero coefficients of a vector w, we consider a penalty involving both a *combinatorial* function F and ℓ_p-regularization:

$$\mathrm{pen} : w \mapsto \mu\, F(\mathrm{Supp}(w)) + \nu\, \|w\|_p^p, \tag{1}$$

where μ and ν are positive scalar coefficients. Since such non-convex discontinuous penalizations are untractable computationally, we undertake to construct an appropriate convex relaxation. The most natural convex surrogate for a non-convex function, say A, is arguably its *convex envelope* (i.e., its tightest convex lower bound) which can be computed as its Fenchel-Legendre bidual A^{**}. However, one relatively natural requirement for a regularizer is to ask that it be also *positively homogeneous* (p.h.) since this leads to formulations that are invariant by rescaling of the data. Our goal will therefore be to construct the tightest positively homogeneous convex lower bound of the penalty considered.

Now, it is a classical result that, given a function A, its tightest p.h. (but not necessarily convex) lower bound A_h is $A_h(w) = \inf_{\lambda>0} \frac{A(\lambda w)}{\lambda}$ (see Rockafellar, 1970, p.35).

This is instrumental here given the following proposition:

Proposition 1. *Let $A : \mathbb{R}^d \to \mathbb{R}_+$ be a real valued function, A_h defined as above. Then C, the tightest positively homogeneous and convex lower bound of A, is well-defined and $C = A_h^{**}$.*

Proof. The set of convex p.h. lower bounds of A is non-empty (since it contains the constant zero function) and stable by taking pointwise maxima. Therefore it has a unique majorant, which we call C. We have for all $w \in \mathbb{R}^d$, $A_h^{**}(w) \leqslant C(w) \leqslant A(w)$, by definition of C and the fact that A_h is an p.h. lower bound on A. We thus have for all $\lambda > 0$, $A_h^{**}(\lambda w)\lambda^{-1} \leqslant C(\lambda w)\lambda^{-1} \leqslant A(\lambda w)\lambda^{-1}$, which implies that for all $w \in \mathbb{R}^d$, $A_h^{**}(w) \leqslant C(w) \leqslant A_h(w)$. Since C is convex, we must have $C = A_h^{**}$, hence the desired result. □

Using its definition we can easily compute the tightest positively homogeneous lower bound of the penalization of Eq. (1), which we denote pen_h:

$$\mathrm{pen}_h(w) \quad = \quad \inf_{\lambda>0} \frac{\mu}{\lambda} F(\mathrm{Supp}(w)) + \nu\, \lambda^{p-1}\, \|w\|_p^p.$$

Setting the gradient of the objective to 0, one gets that the minimum is obtained for $\lambda = \left(\frac{\mu q}{\nu p}\right)^{1/p} F(\mathrm{Supp}(w))^{1/p}\, \|w\|_p^{-1}$, and that

$$\mathrm{pen}_h(w) = (q\mu)^{1/q}\, (p\nu)^{1/p}\, \Theta(w),$$

where we introduced the notation

$$\Theta(w) := \; F(\mathrm{Supp}(w))^{1/q}\, \|w\|_p.$$

Up to a constant factor depending on the choices of μ and ν, we are therefore led to consider the positively homogeneous penalty Θ we just defined, which combines the two terms multiplicatively. Consider the norm Ω_p (or Ω_p^F if a reference to F is needed) whose dual norm[2] is defined as

$$\Omega_p^*(s) := \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}. \tag{2}$$

We have the following result:

Proposition 2 (Convex relaxation)**.** *The norm Ω_p is the convex envelope of Θ.*

Proof. Denote $\Theta(w) = \|w\|_p\, F(\mathrm{Supp}(w))^{1/q}$, and compute its Fenchel conjugate:

$$\begin{aligned}
\Theta^*(s) \quad &= \quad \max_{w\in\mathbb{R}^d} w^\top s - \|w\|_p\, F(\mathrm{Supp}(w))^{1/q} \\
&= \quad \max_{A\subset V} \max_{w_A\in\mathbb{R}_*^{|A|}} w_A^\top s_A - \|w_A\|_p\, F(A)^{1/q} \\
&= \quad \max_{A\subset V} \iota_{\{\|s_A\|_q \leqslant F(A)^{1/q}\}} = \iota_{\{\Omega_p^*(s)\leqslant 1\}},
\end{aligned}$$

[2]The assumptions on the domain $\mathcal{D}_0$ of F and on the positivity of F indeed guarantee that Ω_p^* is a norm.

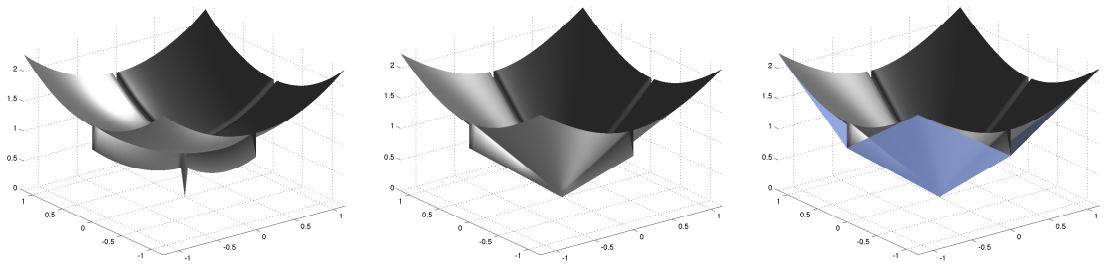

Figure 1: **Penalties in 2D** From left to right: the graph of the penalty $\texttt{pen}$, the graph of penalty $\texttt{pen}_h$ with $p = 2$, and the graph of the norm Ω_2^F in blue overlaid over graph of $\texttt{pen}_h$, for the combinatorial function $F : 2^V \to \mathbb{R}^+$, with $F(\varnothing) = 0$, $F(\{1\}) = F(\{2\}) = 1$ and $F(\{1,2\}) = 1.8$.

where $\iota_{\{s \in S\}}$ is the indicator of the set S, that is the function equal to 0 on S and $+\infty$ on S^c. The Fenchel bidual of Θ, i.e., its largest (thus tightest) convex lower bound, is therefore exactly Ω_p. $\square$

Note that the function F is not assumed submodular in the previous result. Since the function Θ depends on w only through $|w|$, by symmetry, the norm Ω_p is also a function of $|w|$. Given Proposition 1, we have the immediate corollary:

Corollary 1 (Two parts-code relaxation)**.** *Let $p > 1$. The norm $w \mapsto (q\mu)^{1/q}(p\nu)^{1/p}\,\Omega_p(w)$ is the tightest convex* positively homogeneous *lower bound of the function $w \mapsto \mu F(\mathrm{Supp}(w)) + \nu\|w\|_p^p$.*

The penalties and relaxation results considered in this section are illustrated on Figure 1.

2.1 Special cases.

Case $p = 1$. In that case, letting $d_k = \max_{A \ni k} F(A)$, the dual norm is $\Omega_1^*(s) = \max_{k \in V} |s_k|/d_k$ so that $\Omega_1(w) = \sum_{k \in V} d_k\,|w_k|$ is always a weighted ℓ_1-norm. But regularizing with a weighted ℓ_1-norm leads to estimators that can potentially have all sparsity patterns possible (even if some are obviously privileged) and in that sense a weighted ℓ_1-norm cannot encode hard structural constraints on the patterns. Since this means in other words that the ℓ_1-relaxations essentially lose the combinatorial structure of allowed sparsity patterns possibly encoded in F, we focus, from now on, on the case $p > 1$.

Lasso, group Lasso. Ω_p instantiates as the ℓ_1, ℓ_p and ℓ_1/ℓ_p-norms for the simplest functions:

- If $F(A) = |A|$, then $\Omega_p(w) = \|w\|_1$, since $\Omega_p^*(s) = \max_A \frac{\|s_A\|_q}{|A|^{1/q}} = \|s\|_\infty$. It is interesting that the cardinality function is always relaxed to the ℓ_1-norm for all ℓ_p-relaxations, and is not an artifact of the traditional relaxation on an ℓ_∞-ball.
- If $F(A) = 1_{\{A \neq \varnothing\}}$, then $\Omega_p(w) = \|w\|_p$, since $\Omega_p^*(s) = \max_A \|s_A\|_q = \|s\|_q$.
- If $F(A) = \sum_{j=1}^g 1_{\{A \cap G_j \neq \varnothing\}}$, for $(G_j)_{j \in \{1,\dots,g\}}$ a partition of V, then $\Omega_p(w) = \sum_{j=1}^g \|w_{G_j}\|_p$ is the group Lasso or ℓ_1/ℓ_p-norm (Yuan and Lin, 2006). This result provides a principled derivation for the form of these norms, which did not exist in the literature. For groups which do not form a partition, this identity does in fact not hold in general for $p < \infty$, as we discuss in Section 4.

Submodular functions and $p = \infty$. For a submodular function F and in the $p = \infty$ case, the norm Ω^F_∞ that we derived actually coincides with the relaxation proposed by Bach (2010), and as showed in that work, $\Omega^F_\infty(w) = f(|w|)$, where f is a function associated with F and called the *Lovász extension* of F. We discuss the case of submodular functions in detail in Section 6.

2.2 Lower combinatorial envelope

The fact that when F is a submodular function, Ω^F_∞ is equal to the Lovász extension f on the positive orthant provides a guarantee on the tightness of the relaxation. Indeed f is called an "extension" because $\forall A \subset 2^V$, $f(1_A) = F(A)$, so that f can be seen to extend the function F to $\mathbb{R}^d$; as a consequence, $\Omega^F_\infty(1_A) = f(1_A) = F(A)$, which means that the relaxation is tight for all w of the form $w = c\,1_A$, for any scalar constant $c \in \mathbb{R}$ and any set $A \subset V$. If F is not submodular, this property does not necessarily hold, thereby suggesting that the relaxation could be less tight in general. To characterize to which extend this is true, we introduce a couple of new concepts.

Much of the properties of Ω_p, for any $p > 1$, are captured by the unit ball of Ω^*_∞ or its intersection with the positive orthant. In fact, as we will see in the sequel, the ℓ_∞ relaxation plays a particular role, to establish properties of the norm, to construct algorithms and for the statistical analysis, since it it reflects most directly the combinatorial structure of the function F.

We define the *canonical polyhedron*[3] associated to the combinatorial function as the polyhedron $\mathcal{P}_F$ defined by

$$\mathcal{P}_F = \{s \in \mathbb{R}^d,\ \forall A \subset V,\ s(A) \leq F(A)\}.$$

By construction, it is immediate that the unit ball of Ω^*_∞ is $\{s \in \mathbb{R}^d \mid |s| \in \mathcal{P}_F\}$.

From this polyhedron, we construct a new set-function which restitutes the features of F that are captured by $\mathcal{P}_F$:

Definition 2 (Lower combinatorial envelope)**.** *Define the* lower combinatorial envelope *(LCE) of F as the set-function F_- defined by:*

$$F_-(A) = \max_{s \in \mathcal{P}_F} s(A).$$

By construction, even when F is not monotonic, F_- is always non-decreasing (because $\mathcal{P}_F \subset \mathbb{R}^d_+$).

One of the key properties of the lower combinatorial envelope is that, as shown in the next lemma, Ω^F_∞ is an extension of F_- in the same way that the Lovász extension is an extension of F when F is submodular.

Lemma 1. *(Extension property)* $\Omega^F_\infty(1_A) = F_-(A)$.

Proof. From the definition of Ω^F_∞, $\mathcal{P}_F$ and F_-, we get: $\Omega^F_\infty(1_A) = \max_{\Omega^{F\,*}_\infty(s) \leq 1} 1_A^\top s = \max_{s \in \mathcal{P}_F} s^\top 1_A = F_-(A)$ □

Functions that are close to their LCE have in that sense a tighter relaxation than others.

A second important property is that a function F and its LCE share the same canonical polyhedron. This will result as a immediate corollary from the following lemma:

Lemma 2. $\forall s \in \mathbb{R}^V_+$, $\max_{A \subset V} \frac{s(A)}{F(A)} = \max_{A \subset V} \frac{s(A)}{F_-(A)}$.

[3]The reader familiar with submodular functions will recognize that the canonical polyhedron generalizes the submodular polyhedron usually defined for these functions.

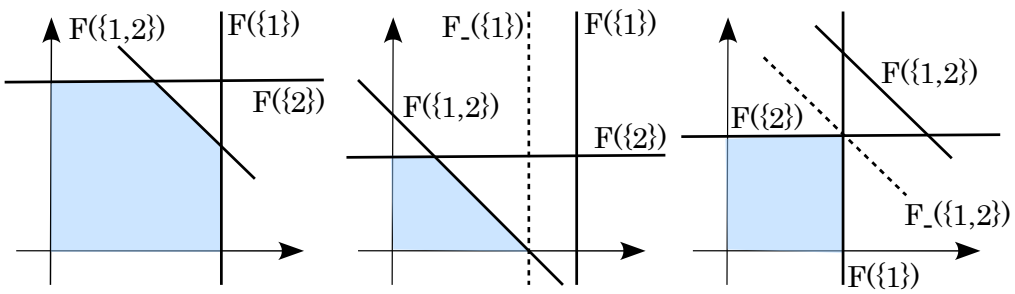

Figure 2: Intersection of the canonical polyhedron with the positive orthant for three different functions F. Full lines materialize the inequalities $s(A) \leq F(A)$ that define the polyhedron. Dashed line materialize the induced constraints $s(A) \leq F_-(A)$ that results from all constraints $s(B) \leq F(B)$, $B \in 2^V$. From left to right: (i) $\mathcal{D}_F = 2^V$ and $F_- = F = F_+$; (ii) $\mathcal{D}_F = \{\{2\}, \{1,2\}\}$ and $F_-(\{1\}) < F(\{1\})$; (iii) $\mathcal{D}_F = \{\{1\}, \{2\}\}$ corresponding to a weighted ℓ_1-norm.

Proof. Given that for all A, $F_-(A) \leqslant F(A)$, the left hand side is always smaller or equal to the right hand side. We now reason by contradiction. Assume that there exists s such that $\forall A \subset V$, $s(A) \leq \nu F(A)$ but that there exists $B \subset V$, $s(B) > \nu F(B)$, then $s' = \frac{1}{\nu} s$ satisfies $\forall A \subset V$, $s'(A) \leq F(A)$. By definition of F_-, the latter implies that $F_-(B) \geq s'(B) = \frac{1}{\nu} s(B) > \frac{1}{\nu} \cdot \nu F(B)$, where the last inequality results from the choice of this particular B. This would imply $F_-(B) > F(B)$, but by definition of F_-, we have $F_-(A) \leq F(A)$ for all $A \subset V$. □

Corollary 3. $\mathcal{P}_F = \mathcal{P}_{F_-}$.

But the sets $\{w \in \mathbb{R}^d \mid |w| \in \mathcal{P}_F\}$ and $\{w \in \mathbb{R}^d \mid |w| \in \mathcal{P}_{F_-}\}$ are respectively the unit balls of Ω_∞^F and $\Omega_\infty^{F_-}$. As a direct consequence, we have:

Lemma 3. *For all* $p \geq 1$, $\quad \Omega_p^F = \Omega_p^{F_-}$.

By construction, F_- is the largest function which lower bounds F, and has the same ℓ_p-relaxation as F, hence the term of *lower combinatorial envelope*.

Figure 2 illustrates the fact that F and F_- share the same canonical polyhedron and that the value of $F_-(A)$ is determined by the values that F takes on other sets. This figure also suggests that some constraints $\{\, s(A) \leq F(A) \,\}$ can never be active and could therefore be removed. This will be formalized in Section 2.3.

To illustrate the relevance of the concept of lower combinatorial envelope, we compute it for a specify combinatorial function, the range function, and show that it enables us to answer the question of whether the relaxation would be good in this case.

Example 1 (Range function)**.** *Consider, on* $V = [1, d]$*, the range function* $F : A \mapsto \max(A) - \min(A) + 1$ *where* $\min(A)$ *(resp.* $\max(A)$*) is the smallest (resp. largest) element in* A*. A motivation to consider this function is that it induces the selection of supports that are exactly intervals. Since* $F(\{i\}) = 1$*,* $i \in V$*, then for all* $s \in \mathcal{P}_F$*, we have* $s(A) \leq |A| \leq F(A)$*. But this implies that* $\mathcal{D}_F$ *is the set of singletons and that* $F_-(A) = |A|$*, so that* Ω^F *is the* ℓ_1*-norm and is oblivious of the structure encoded in* F*.*

As we see from this example, the lower combinatorial envelope can be interpreted as the combinatorial function which the relaxation is actually able to capture.

2.3 Upper combinatorial envelope

Let F be a set-function and $\mathcal{P}_F$ its canonical polyhedron. In this section, we follow an intuition conveyed by Figure 2 and find a compact representation of F: the polyhedron $\mathcal{P}_F$ has in many cases a number of faces which much smaller than 2^d. We formalize this in the next lemma.

Lemma 4. *There exists a minimal subset $\mathcal{D}_F$ of 2^V such that for $s \in \mathbb{R}^d_+$,*

$$s \in \mathcal{P}_F \Leftrightarrow (\forall A \in \mathcal{D}_F,\ s(A) \leq F(A)).$$

Proof. To prove the result, we define as in Obozinski et al. (2011, Sec. 8.1) the notion of *redundant* sets: we say that a set A is redundant for F if

$$\exists A_1, \dots, A_k \in 2^V \backslash \{A\}, \quad (\forall i,\ s(A_i) \leq F(A_i)) \Rightarrow (s(A) \leq F(A)).$$

Consider the set $\mathcal{D}_F$ of all non redundant sets.

We will show that, in fact, A is redundant for F if and only if

$$\exists A_1, \dots, A_k \in \mathcal{D}_F \backslash \{A\}, \quad (\forall i,\ s(A_i) \leq F(A_i)) \Rightarrow (s(A) \leq F(A)),$$

which proves the lemma.

Indeed, we can use a peeling argument to remove all redundant sets one by one and show recursively that the inequality constraint associated with a given redundant set is still implied by all the ones we have not removed yet. The procedure stops when we have reached the smallest set $\mathcal{D}_F$ of constraints implying all the other ones. □

We call $\mathcal{D}_F$ the *core set* of F. It corresponds to the set of faces of dimension $d-1$ of $\mathcal{P}_F$.

This notion motivates the definition of a new set-function:

Definition 4. *(Upper combinatorial envelope) We call* upper combinatorial envelope *(UCE) the function F_+ defined by $F_+(A) = F(A)$ for $A \in \mathcal{D}_F$ and $F_+(A) = \infty$ otherwise.*

As the reader might expect at this point, F_+ provides a compact representation which captures all the information about F that is preserved in the relaxation:

Proposition 3. *F, F_- and F_+ all define the same canonical polyhedron $\mathcal{P}_{F_-} = \mathcal{P}_F = \mathcal{P}_{F_+}$ and share the same core set $\mathcal{D}_F$. Moreover, $\forall A \in \mathcal{D}_F$, $F_-(A) = F(A) = F_+(A)$.*

Proof. To show that $\Omega_p^{F_+} = \Omega_p^F$ we just need to show $\mathcal{P}_{F_+} = \mathcal{P}_F$. By the definition of F_+ we have $\mathcal{P}_{F_+} = \{s \in \mathbb{R}^d \mid s(A) \leq F(A),\ A \in \mathcal{D}_F\}$ but the previous lemma precisely states that the last set is equal to $\mathcal{P}_F$.

We now argue that, for all $A \in \mathcal{D}_F$, $F_-(A) = F(A) = F_+(A)$. Indeed, the equality $F(A) = F_+(A)$ holds by definition, and, for all $A \in \mathcal{D}_F$, we need to have $F(A) = F_-(A)$ because $F_-(A) = \max_{s \in \mathcal{P}_F} s(A) = \max_{s \in \mathcal{P}_{F_+}} s(A)$ and if we had $F_-(A) < F(A)$, this would imply that A is redundant. □

Finally, the term "upper combinatorial envelope" is motivated by the following lemma:

Lemma 5. *F_+ is the pointwise supremum of all the set-functions H that are upper bounds on F and such that $\mathcal{P}_H = \mathcal{P}_F$.*

Proof. We need to show that we have $F_+ : A \mapsto \sup\{H(A) \mid H \in \mathcal{E}_F\}$ with

$$\mathcal{E}_F := \{H : 2^V \to \overline{\mathbb{R}}_+ \text{ s.t. } H \geq F \text{ and } \mathcal{P}_H = \mathcal{P}_F\}.$$

But for any $H \in \mathcal{E}_F$, $\mathcal{P}_H = \mathcal{P}_F$ implies that $H_- = F_-$ by definition of the lower combinatorial envelope. Moreover we have $\mathcal{D}_H \subset \mathcal{D}_F$ since if $A \notin \mathcal{D}_F$, then A is redundant for F, i.e. there are $A_1, \dots, A_k \in \mathcal{D}_F$ such that $\big(\forall i,\ s(A_i) \leq F_-(A_i)\big) \Rightarrow (s(A) \leq F(A))$, but $F(A) \leq H(A)$, and thus, since $F_-(A_i) = H_-(A_i)$, this implies that A is redundant for H, i.e., $A \notin \mathcal{D}_H$. Now, assume there is $A \in \mathcal{D}_F \cap \mathcal{D}_H^c$. Since A is redundant for H then there are $A_1, \dots, A_k$ in $\mathcal{D}_H \backslash \{A\}$ such that $\big(\forall i,\ s(A_i) \leq H(A_i)\big) \Rightarrow (s(A) \leq H(A))$, but since $A_i \in \mathcal{D}_H \subset \mathcal{D}_F$ we have $H(A_i) = H_-(A_i) = F_-(A_i)$ and since $A \in \mathcal{D}_F$ we also have $F(A) = F_-(A)$ so either $\big(\forall A' \in \mathcal{D}_H,\ s(A') \leq F_-(A')\big) \Rightarrow (s(A) \leq F_-(A))$, so that A is redundant, which is excluded, or since $H_-(A) = \max_{s \in \mathcal{P}_H} s(A) = \max_{s:\, s(A') \leq F_-(A'),\ A' \in \mathcal{D}_H} s(A)$, we then have $H_-(A) > F_-(A)$, which is also impossible. So we necessarily have $\mathcal{D}_H = \mathcal{D}_F$. To conclude the proof we just need to show that $F_+ \in \mathcal{E}_F$ and that $F_+ \geq H$ for all $H \in \mathcal{E}_F$; this inequality is trivially satisfied for $A \notin \mathcal{D}_{F_+}$, and since $\mathcal{D}_H = \mathcal{D}_{F_+}$, for $A \in \mathcal{D}_{F_+}$, we have $F_+(A) = F_-(A) = H_-(A) = H(A)$. □

The picture that emerges at this point from the results shown is rather simple: any combinatorial function F defines a polyhedron $\mathcal{P}_F$ whose faces of dimension $d-1$ are indexed by a set $\mathcal{D}_F \subset 2^V$ that we called the *core set*. In symbolic notation: $\mathcal{P}_F = \{s \in \mathbb{R}^d \mid s(A) \leq F(A),\ A \in \mathcal{D}_F\}$. All the combinatorial functions which are equal to F on $\mathcal{D}_F$ and which otherwise take values that are larger than its lower combinatorial envelope F_-, have the same ℓ_p tightest positively homogeneous convex relaxation Ω_p^F, the smallest such function being F_- and the largest F_+. Moreover $F_-(A) = \Omega_\infty^F(\mathrm{A})$, so that Ω_∞^F is an extension of F_-. By construction, and even if F is a non-decreasing function, F_- is non-decreasing, while F_+ is obviously not a decreasing function, even though its restriction to $\mathcal{D}_F$ is. It might therefore seem an odd set-function to consider; however if $\mathcal{D}_F$ is a small set, since $\Omega_p^F = \Omega_p^{F_+}$, and it provides a potentially much more compact representation of the norm, which we now relate to a norm previously introduced in the literature.

3 Latent group Lasso, block-coding and set-cover penalties

The norm Ω_p is actually not a new norm. It was introduced from a different point of view by Jacob et al. (2009) (see also Obozinski et al., 2011) as one of the possible generalizations of the group Lasso to the case where groups overlap.

To establish the connection, we now provide a more explicit form for Ω_p, which is different from the definition via its dual norm which we have exploited so far.

We consider models that are parameterized by a vector $w \in \mathbb{R}^V$ and associate to them latent variables that are tuples of vectors of $\mathbb{R}^V$ indexed by the power-set of V. Precisely, with the notation

$$\mathcal{V} = \big\{v = (v^A)_{A \subset V} \in \big(\mathbb{R}^V\big)^{2^V} \text{ s.t. } \mathrm{Supp}(v^A) \subset A\big\},$$

we define the norms Ω_p as

$$\Omega_p(w) = \min_{v \in \mathcal{V}} \sum_{A \subset V} F(A)^{\frac{1}{q}} \|v^A\|_p \text{ s.t. } w = \sum_{A \subset V} v^A. \tag{3}$$

As suggested by notations and as first proved for $p = 2$ by Jacob et al. (2009), we have:

Lemma 6. *Ω_p and Ω_p^* are dual to each other.*

An elementary proof of this result is provided in Obozinski et al. (2011)[4]. We propose a slightly more abstract proof of this result in appendix A using explicitly the fact that Ω_p is defined as an infimal convolution.

We will refer to this norm Ω_p as the *latent group Lasso* since it is defined by introducing latent variables v^A that are themselves regularized instead of the original model parameters. We refer the reader to Obozinski et al. (2011) for a detailed presentation of this norm, some of its properties and some support recovery results in terms of the support of the latent variables. In Jacob et al. (2009) the expansion (3) did not involve all terms of the power-set but only a subcollection of sets $\mathcal{G} \subset 2^V$. The notion of redundant set discussed in Section 2.3 was actually introduced by Obozinski et al. (2011, Sec. 8.1) and the set $\mathcal{G}$ could be viewed as the core set $\mathcal{D}_F$. A result of Obozinski et al. (2011) for $p = 2$ generalizes immediately to other p: the unit ball of Ω_p can be shown to be the convex hull of the sets $D_A = \{w \in \mathbb{R}^d \mid \|w_A\|_p^p \leq F(A)^{-1/q}\}$. This is illustrated in Figure 3.

The motivation of Jacob et al. (2009) was to find a convex regularization which would induce sparsity patterns that are unions of groups in $\mathcal{G}$ and explain the estimated vector w as a combination of a small number of latent components, each supported on one group of $\mathcal{G}$. The motivation is very similar in Huang et al. (2011) who consider an ℓ_0-type penalty they call *block coding*, where each support is penalized by the minimal sum of the *coding complexities* of a certain number of elementary sets called "blocks" which *cover* the support. In both cases the underlying combinatorial penalty is the *minimal weighted set cover* defined for a set $B \subset V$ by:

$$\widetilde{F}(B) \;=\; \min_{(\delta^A)_{A\subset V}} \sum_{A\subset V} F(A)\,\delta^A \qquad \text{s.t.} \qquad \sum_{A\subset V} \delta^A 1_A \geq 1_B, \qquad \delta^A \in \{0,1\},\ A \subset V.$$

While the norm proposed by Jacob et al. (2009) can be viewed as a form of "relaxation" of the cover-set problem, a rigorous link between the ℓ_0 and convex formulation is missing. We will make this statement rigorous through a new interpretation of the lower combinatorial envelope of F.

Indeed, assume w.l.o.g. that $w \in \mathbb{R}^d_+$. For $x, y \in \mathbb{R}^V$, we write $x \geq y$ if $x_i \geq y_i$ for all $i \in V$. Then,

$$\begin{aligned}
\Omega_\infty(w) \;&=\; \min_{v\in\mathcal{V}} \sum_{A\subset V} F(A)\|v^A\|_\infty \qquad \text{s.t.} \qquad \sum_{A\subset V} v^A \geq w \\
&=\; \min_{\delta^A\in\mathbb{R}_+} \sum_{A\subset V} F(A)\,\delta^A \qquad \text{s.t.} \qquad \sum_{A\subset V} \delta^A 1_A \geq w,
\end{aligned}$$

since if $(v^A)_{A\subset V}$ is a solution so is $(\delta^A 1_A)_{A\subset V}$ with $\delta^A = \|v^A\|_\infty$. We then have

$$F_-(B) \;=\; \min_{(\delta^A)} \sum_{A\subset V} F(A)\,\delta^A, \qquad \text{s.t.} \qquad \sum_{A\subset V} \delta^A 1_A \geq 1_B, \quad \delta^A \in [0,1],\ A \subset V, \tag{4}$$

because constraining δ to the unit cube does not change the optimal solution, given that $1_B \leq 1$. But the optimization problem in (4) is exactly the *fractional weighted set-cover problem* (Lovász, 1975), a classical relaxation of the *weighted cover set problem* in Eq. (4).

Combining Proposition 2 with the fact that $F_-(A)$ is the fractional weighted set-cover, now yields:

Theorem 5. *$\Omega_p(w)$ is the tightest convex relaxation of the function $w \mapsto \|w\|_p\, \widetilde{F}(\mathrm{Supp}(w))^{1/q}$ where $\widetilde{F}(\mathrm{Supp}(w))$ is the* weighted set-cover *of the support of w.*

Proof. We have $F_-(A) \leq \widetilde{F}(A) \leq F(A)$ so that, since F_- is the lower combinatorial envelope of F, it is also the lower combinatorial envelope of $\widetilde{F}$, and therefore $\Omega_p^{F_-} = \Omega_p^{\widetilde{F}} = \Omega_p^F$. □

[4]The proof in Obozinski et al. (2011) addresses the $p = 2$ case but generalizes immediately to other values of p.

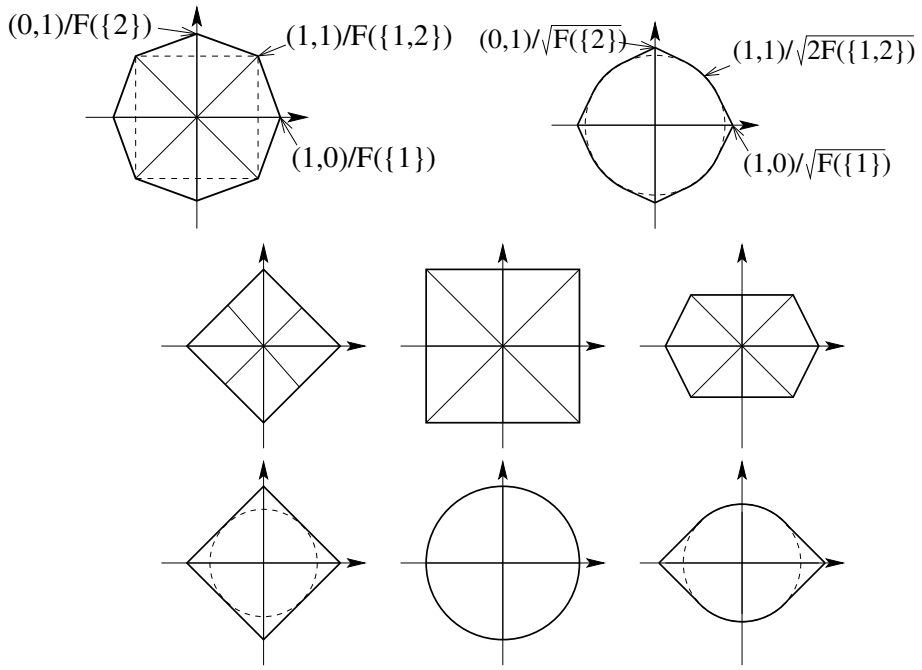

Figure 3: Unit balls in $\mathbb{R}^2$ for four combinatorial functions (actually all submodular) on two variables. Top left and middle row: $p = \infty$; top right and bottom row: $p = 2$. Changing values of F may make some of the extreme points disappear. All norms are hulls of a disk and points along the axes, whose size and position is determined by the values taken by F. On top row: $F(A) = F_-(A) = |A|^{1/2}$ (all possible extreme points); and from left to right on the middle and bottom rows: $F(A) = |A|$ (leading to $\|\cdot\|_1$), $F(A) = F_-(A) = \min\{|A|, 1\}$ (leading to $\|\cdot\|_p$), $F(A) = F_-(A) = \frac{1}{2}1_{\{A\cap\{2\}\neq\varnothing\}} + 1_{\{A\neq\varnothing\}}$.

This proves that the norm Ω_p^F proposed by Jacob et al. (2009) is indeed in a rigorous sense a relaxation of the block-coding or set-cover penalty.

Example 2. *To illustrate the above results consider the block-coding scheme for subsets of* $V = \{1, 2, 3\}$ *with blocks consisting only of pairs, i.e., chosen from the collection* $\mathcal{D}_0 := \big\{\{1, 2\}, \{2, 3\}, \{1, 3\}\big\}$ *with costs all equal to 1. The following table lists the values of* F, F_- *and* $\widetilde{F}$*:*

	$\varnothing$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{2,3\}$	$\{1,3\}$	$\{1,2,3\}$
F	0	∞	∞	∞	1	1	1	∞
$\widetilde{F}$	0	1	1	1	1	1	1	2
F_-	0	1	1	1	1	1	1	3/2

Here, F *is equal to its UCE (except that* $F_+(\varnothing) = \infty$*) and takes therefore non trivial values only on the core set* $\mathcal{D}_F = \mathcal{D}_0$*. All non-empty sets except* V *can be covered by exactly one set, which explains the cases where* F_- *and* $\widetilde{F}$ *take the value one.* $\widetilde{F}(V) = 2$ *since* V *is covered by any pair of blocks and a slight improvement is obtained if fractional covers is allowed since for* $\delta_1 = \delta_2 = \delta_3 = \frac{1}{2}$*, we have* $1_V = \delta_1 1_{\{2,3\}} + \delta_2 1_{\{3,1\}} + \delta_3 1_{\{1,2\}}$ *and therefore* $F_-(V) = \delta_1 + \delta_2 + \delta_3 = \frac{3}{2}$*.*

The interpretation of the LCE as the value of a minimum fractional weighted set cover suggests a new interpretation of F_+ (or equivalently of $\mathcal{D}_F$) as defining the smallest set of blocks ($\mathcal{D}_F$) and their costs, that induce a fractional set over problem with the same optimal value.

It is interesting to note (but probably not a coincidence) that it is Lovász who introduced the concept of optimal fractional weighted set cover, while we just showed that the value of that cover is precisely F_-, i.e., the combinatorial function which is extended by $\Omega_\infty^{F_+} = \Omega_\infty^{F_-}$ and which, if F_+ is submodular is equal to the Lovász extension.

The interpretation of F_- as the value of a minimum fractional weighted cover set problem allows us also to show a result which is dual to the property of LCEs, and which we now present.

3.1 Largest convex positively homogeneous function with same combinatorial restriction

By symmetry with the characterization of the *lower combinatorial envelope* as the smallest combinatorial function that has the same tightest convex and positively homogeneous (p.h.) relaxation as a given combinatorial function F, we can, given a convex positively homogeneous function g, define the combinatorial function $F : A \mapsto g(1_A)$, which by construction, is the combinatorial function which g *extends* (in the sense of Lovász) to $\mathbb{R}_+^d$, and ask if there exists a largest convex and p.h. function g^+ among all such functions. It turns out that this problem is well-posed if the question is restricted to functions that are also coordinate-wise non-decreasing. Perhaps not surprisingly, it is then the case that the largest convex p.h. function extending the same induced combinatorial function is precisely Ω_∞^F, as we show in the next lemma.

Lemma 7. *(Largest convex positively homogeneous extension) Let g be a convex, p.h. and coordinate-wise non-decreasing function defined on $\mathbb{R}_+^d$. Define F as $F : A \mapsto g(1_A)$ and denote by F_- its lower combinatorial envelope.*
Then $F = F_-$ and $\forall w \in \mathbb{R}^d$, $g(|w|) \leq \Omega_\infty^F(w)$.

Proof. From Equation (4), we know that F_- can be written as the value of a minimal weighted fractional set-cover. But if $1_B \leq \sum_{A \subset V} \delta^A 1_A$, we have

$$\sum_{A \subset V} \delta^A g(1_A) \geq g\big(\textstyle\sum_{A \subset V} \delta^A\big) \geq g(1_B),$$

where the first inequality results from the convexity and homogeneity of g, and the second from the assumption that it is coordinate-wise non-decreasing. As a consequence, injecting the above inequality in (4), we have $F_-(B) \geq F(B)$. But since, we always have $F_- \leq F$, this proves the equality.

For the second statement, using the coordinate-wise monotonicity of g and its homogeneity, we have $g(|w|) \leq \|w\|_\infty g(1_{\mathrm{Supp}(w)}) = \|w\|_\infty F(\mathrm{Supp}(w))$. Then, taking the convex envelope of functions on both sides of the inequality we get $g(|\cdot|)^{**} \leq \big(\|\cdot\|_\infty F(\mathrm{Supp}(\cdot))\big)^{**} = \Omega_\infty^F$, where $(\cdot)^*$ denotes the Fenchel-Legendre transform. □

4 Examples

subsectionOverlap count functions, their relaxations and the ℓ_1/ℓ_p-norms. A natural family of set functions to consider are the functions that, given a collection of sets $\mathcal{G} \subset 2^V$ are defined as the number of these sets that are intersected by the support:

$$F_\cap(A) = \sum_{B \in \mathcal{G}} d_B 1_{\{A \cap G \neq \varnothing\}}. \tag{5}$$

Since $A \mapsto 1_{\{A \cap G \neq \varnothing\}}$ is clearly submodular and since submodular functions form a positive cone, all these functions are submodular, which implies that $\Omega_p^{F_\cap}$ is a tight relaxation of $F_\cap$.

Overlap count functions *vs* set-covers. As mentioned in Section 2.1, if $\mathcal{G}$ is a partition, the norm $\Omega_p^{F_\cap}$ is the ℓ_1/ℓ_p-norm; in this special case, $F_\cap$ is actually the value of the minimal (integer-valued) weighted set-cover associated with the sets in $\mathcal{G}$ and the weights d_G.

However, it should be noted that, in general, the value of these functions is quite different from the value of a minimal weighted set-cover. It has rather the flavor of some sort of "maximal weighted set-cover" in the sense that any set that has a non-empty intersection in the support would be included in the cover. We call them overlap count functions.

ℓ_p relaxations of $F_\cap$ *vs* ℓ_1/ℓ_p-norms. In the case where $p = \infty$, Bach (2010) showed that even when groups overlap we have $\Omega_\infty(w) = \sum_{B \in \mathcal{G}} d_B \|w_G\|_\infty$, since the Lovász extension of a sum of submodular functions is just the sum of the Lovász extensions of the terms in the sum.

The situation is more subtle when $p < \infty$: in that case, and perhaps surprisingly, $\Omega_p^{F_\cap}$ is not the *weighted ℓ_1/ℓ_p norm with overlap* (Jenatton et al., 2011a), also referred to as the *overlapping group Lasso* (which should clearly be distinguished from the *latent group Lasso*) and which is the norm defined by $w \mapsto \sum_{B \in \mathcal{G}} d'_B \|w_G\|_p$. The norm $\Omega_p^{F_\cap}$ does not have a simple closed form in general. In terms of sparsity patterns induced however, $\Omega_p^{F_\cap}$ behaves like $\Omega_\infty^{F_\cap}$, and as a result the sparsity patterns allowed by $\Omega_p^{F_\cap}$ are the same as those allowed by the corresponding *weighted ℓ_1/ℓ_p norm with overlap*.

ℓ_p-relaxation of $F_\cap$ *vs* latent group Lasso based on $\mathcal{G}$. It should be clear as well that $\Omega_p^{F_\cap}$ is not itself the *latent group Lasso* associated with the collection $\mathcal{G}$ and the weights d_G in the sense of Jacob et al. (2009). Indeed, the latter corresponds to the function $F_\cup : A \mapsto 1_{\{A \neq \varnothing\}} + \iota_{\{A \in \mathcal{G}\}}$, or to its LCE which is the minimal value of the fractional weighted set cover associated with $\mathcal{G}$. Clearly, $F_\cup$ is in general strictly smaller than $F_\cap$ and since the relaxation of the latter is tight, it cannot be equal to the relaxation of the former, if the combinatorial functions are themselves different. Obviously, the function $\Omega_p^{F_\cap}$ is still as shown in this paper, another latent group Lasso corresponding to a fractional weighted set cover and involving a larger number of sets that the ones in $\mathcal{G}$ (possibly all of 2^V). This last statement leads us to what might appear to be a paradox, which we discuss next.

Supports *stable by intersection* vs *formed as unions*. Jenatton et al. (2011a) have shown that the family of norms they considered induces possible supports which form a family that is *stable by intersection*, in the sense that the intersection of any two possible support is also a possible support. But since as mentioned above they have the same support as the norms $\Omega_p^{F_\cap}$, for $1 < p \leq \infty$, which are latent group Lasso norms, and since Jacob et al. (2009) have discussed the fact that the supports induced by any norm Ω_p are formed by *unions* of elements of the core set $\mathcal{D}$, is might appear paradoxical that the allowed support can be described at the same time as intersections and as unions. There is in fact not contradiction because in general the set of supports that are induced by the latent group Lasso are in fact not *stable by union* in the sense that some unions are actually "unstable" and will thus not be selected.

Three different norms. To conclude, we must, given a set of groups $\mathcal{G}$ and a collection of weights $(d_G)_{G \in \mathcal{G}}$, distinguish three norms that can be defined from it, the weighted ℓ_1/ℓ_p-norm with overlap,

the norm $\Omega_p^{F_\cap}$ obtained as the ℓ_p relaxation of the submodular penalty $F_\cap$, and finally, the norm $\Omega_p^{F_{[\cup]}}$ obtained as the relaxation of the set-cover or block-coding penalty with the weights d_G.

Some of the advantages of using a tight relaxation still need to be assessed empirically and theoretically, but the possibility of using ℓ_p-relaxation for $p < \infty$ removes the artifacts that were specific to the ℓ_∞ case.

4.1 Chains, trees and directed acyclic graphs.

Instances of the three types of norms above are naturally relevant to induce sparsity pattern on structures such as chains, trees and directed acyclic graphs.

The weighted ℓ_1/ℓ_p-norm with overlap has been proposed to induce interval patterns on chains and rectangular or convex patterns on grids (Jenatton et al., 2011a), for certain sparsity patterns on trees (Jenatton et al., 2011b) and on directed acyclic graphs (Mairal et al., 2011).

One of the norm considered in Jenatton et al. (2011a) provides a nice example of an overlap count function, which it is worth presenting.

Example 3 (Modified range function)**.** *A shown in Example 1 in Section 2.2, the natural range function on a sequence leads to a trivial LCE. Consider now the penalty with the form of Eq. (5) with $\mathcal{G}$ the set of groups defined as*

$$\mathcal{G} = \big\{[\![1, k]\!] \mid 1 \leq k \leq p\big\} \cup \big\{[\![k, p]\!] \mid 1 \leq k \leq p\big\}.$$

A simple calculation shows that $F_\cap(\varnothing) = 0$ and that for $A \neq \varnothing$, $F_\cap(A) = d - 1 + range(A)$. This function is submodular as a sum of submodular functions, and thus equal to it lower combinatorial envelope, which implies that the relaxation retains the structural a prior encoded by the combinatorial function itself. We will consider the ℓ_2 relaxation of this submodular function in the experiments (see Section 7) and compare it with the ℓ_1/ℓ_2-norm with overlap of Jenatton et al. (2011a).

In the case of trees and DAGs, a natural counting function to consider is the number of nodes which have at least one descendant in the support, i.e. functions of the form $F_\cap : A \mapsto \sum_{i \in V} 1_{\{A \cap D_i \neq \varnothing\}}$, where D_i is the set containing node i and all its descendants. It is related to the weighted ℓ_1/ℓ_p-norms which were considered in Jenatton et al. (2011b) ($p \in \{2, \infty\}$) for and Mairal and Yu (2012) ($p = \infty$). As discussed before, while these norms include $\Omega_\infty^{F_\cap}$ if $p = \infty$, they otherwise do not correspond to the tightest relaxation, which it would be interesting to consider in future work.

Beyond the standard group Lasso and the exclusive group Lasso, there are very few instances of the norm Ω_2^F appearing in the literature. One such example is the wedge penalty considered in Micchelli et al. (2011).

Latent group Lasso formulations are also of interest in these cases, and have not been yet been investigated much, with the exception of Mairal and Yu (2012), which considered the case of a parameter vector with coefficients indexed by a DAG and $\mathcal{G}$ the set of all paths in the graph.

There are clearly other combinatorial functions of interest than submodular functions and set-cover functions. We present an example of such functions in the next section.

4.2 Exclusive Lasso

The exclusive Lasso is a formulation proposed by Zhou et al. (2010) which considers the case where a partition $\mathcal{G} = \{G_1, \ldots, G_k\}$ of V is given and the sparsity imposed is that w should have at most

one non-zero coefficient in each group G_j. The regularizer proposed by Zhou et al. (2010) is the ℓ_p/ℓ_1-norm defined[5] by $\|w\|_{\ell_p/\ell_1} = (\sum_{G\in\mathcal{G}} \|w_G\|_1^p)^{1/p}$. Is this the tightest relaxation?

A natural combinatorial function corresponding to the desired constraint is the function $F(A)$ defined by $F(\varnothing) = 0$, $F(A) = 1$ if $\max_{G\in\mathcal{G}} |A \cap G| = 1$ and $F(A) = \infty$ otherwise.

To characterize the corresponding Ω_p we can compute explicitly its dual norm Ω_p^*:

$$\begin{aligned}
\big(\Omega_p^*(w)\big)^q &= \max_{A\subset V,\, A\neq\varnothing} \frac{\|s_A\|_q^q}{F(A)} \\
&= \max_{A\subset V} \|s_A\|_q^q \quad \text{s.t.} \quad |A\cap G| \leq 1,\, G\in\mathcal{G} \\
&= \max_{i_j\in G_j,\, 1\leq j\leq k} \sum_{j=1}^k |s_{i_j}|^q \;=\; \sum_{j=1}^k \max_{i\in G_j} |s_{i_j}|^q \;=\; \sum_{j=1}^k \|s_{G_j}\|_\infty^q,
\end{aligned}$$

which shows that Ω_p^* is the ℓ_q/ℓ_∞-norm or equivalently that Ω_p is the ℓ_p/ℓ_1-norm and provides a theoretical justification for the choice of this norm: it is indeed the tightest relaxation! It is interesting to compute the lower combinatorial extension of F which is $F_-(A) = \Omega_\infty^F(1_A) = \|1_A\|_{\ell_\infty/\ell_1} = \max_{G\in\mathcal{G}} |A\cap G|$. This last function is also a natural combinatorial function to consider; by the previous result F_- has the same convex relaxation as F, but it would be however less obvious to show directly that $\Omega_p^{F_-}$ is the ℓ_p/ℓ_1 (see appendix B for a direct proof which uses Lemma 7).

5 A variational form of the norm

Several results on Ω_p rely on the fact that it can be related variationally to Ω_∞.

Lemma 8. *Ω_p admits the two following variational formulations:*

$$\begin{aligned}
\Omega_p(w) &= \max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V} \kappa_i^{1/q} |w_i| \quad \textit{s.t.} \quad \forall A\subset V,\, \kappa(A) \leq F(A) \\
&= \min_{\eta\in\mathbb{R}_+^d} \sum_{i\in V} \frac{1}{p} \frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\Omega_\infty(\eta).
\end{aligned}$$

Proof. Using Fenchel duality, we have:

$$\begin{aligned}
\Omega_p(w) &= \max_{s\in\mathbb{R}^d} s^\top w \quad \text{s.t.} \quad \Omega_p^*(w) \leq 1 \\
&= \max_{s\in\mathbb{R}^d} s^\top w \quad \text{s.t.} \quad \forall A\subset V,\, \|s_A\|_q^q \leq F(A) \text{ by definition of } \Omega_p^*, \\
&= \max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V} \kappa_i^{1/q} |w_i| \quad \text{s.t.} \quad \forall A\subset V,\, \kappa(A) \leq F(A).
\end{aligned}$$

[5]The Exclusive Lasso norm which is ℓ_p/ℓ_1 should not be confused with the group Lasso norm which is ℓ_1/ℓ_p.

But it is easy to verify that $\kappa_i^{1/q}|w_i| = \min_{\eta_i \in \mathbb{R}_+} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta_i\kappa_i$ with the minimum attained for $\eta_i = \frac{|w_i|}{\kappa_i^{1/p}}$. We therefore get:

$$\begin{aligned}
\Omega_p(w) &= \max_{\kappa \in \mathbb{R}_+^d} \min_{\eta \in \mathbb{R}_+^d} \sum_{i \in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta^\top \kappa \quad \text{s.t.} \quad \forall A \subset V,\ \kappa(A) \leq F(A) \\
&= \min_{\eta \in \mathbb{R}_+^d} \max_{\kappa \in \mathbb{R}_+^d} \sum_{i \in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\eta^\top \kappa \quad \text{s.t.} \quad \forall A \subset V,\ \kappa(A) \leq F(A) \\
&= \min_{\eta \in \mathbb{R}_+^d} \sum_{i \in V} \frac{1}{p}\frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q}\Omega_\infty(\eta),
\end{aligned}$$

where we could exchange minimization and maximization since the function is convex-concave in η and κ, and where we eliminated formally κ by introducing the value of the dual norm $\Omega_\infty(\eta) = \max_{\kappa \in \mathcal{P}_F} \kappa^\top \eta$. □

Since Ω_∞ is convex, the last formulation is actually jointly convex in (w, η) since $(x, z) \mapsto \frac{1}{p}\frac{\|x\|_p^p}{z^{p-1}} + \frac{1}{q}z$ is convex, as the perspective function of $t \mapsto t^p$ (see Boyd and Vandenberghe, 2004, p. 89).

It should be noted that the norms Ω_p therefore belong to the broad family of H-norms as defined[6] in Bach et al. (2012, Sec. 1.4.2.) and studied by Micchelli et al. (2011).

The above result is particularly interesting if F is submodular since Ω_∞ is then equal to the Lovász extension of F on the positive orthant (Bach, 2010). In this case in particular, it is possible, as we will see in the next section to propose efficient algorithms to compute Ω_p and Ω_p^*, the associated proximal operators, and algorithms to solve learning problems regularized with Ω_p thanks to the above variational form.

For submodular functions, these variational forms are also the basis for the *local decomposability* result of Section 6.4 which is key to establish support recovery in Section 6.5.

6 The case of submodular penalties

In this section, we focus on the case where the combinatorial function F is submodular.

Specifically, we will consider a function F defined on the power set 2^V of $V = \{1, \dots, d\}$, which is *nondecreasing* and *submodular*, meaning that it satisfies respectively

$$\forall A, B \subset V, \qquad A \subset B \Rightarrow F(A) \leqslant F(B),$$

Moreover, we assume that $F(\varnothing) = 0$. These set-functions are often referred to as *polymatroid set-functions* (Fujishige, 2005; Edmonds, 2003). Also, without loss of generality, we assume that F is strictly positive on singletons, i.e., for all $k \in V$, $F(\{k\}) > 0$. Indeed, if $F(\{k\}) = 0$, then by submodularity and monotonicity, if $A \ni k$, $F(A) = F(A \backslash \{k\})$ and thus we can simply consider $V \backslash \{k\}$ instead of V.

Classical examples are the cardinality function and, given a partition of V into $G_1 \cup \cdots \cup G_k = V$, the set-function $A \mapsto F(A)$ which is equal to the number of groups $G_1, \dots, G_k$ with non empty intersection with A, which, as mentioned in section 2.1 leads to the grouped ℓ_1/ℓ_p-norm.

[6] Note that H-norms are in these references defined for $p = 2$ and that the variational formulation proposed here generalizes this to other values of $p \in (1, \infty)$

With a slightly different perspective than the approach of this paper, Bach (2010) studied the special case of the norm Ω_p^F when $p = \infty$ and F is submodular. As mentioned previously, he showed that in that case the norm Ω_∞^F is the Lovász extension of the submodular function F, which is a well studied mathematical object.

Before presenting results on ℓ_p relaxations of submodular penalties, we review a certain number of relevant properties and concepts from submodular analysis. For more details, see, e.g., Fujishige (2005), and, for a review with proofs derived from classical convex analysis, see, e.g., Bach (2011).

6.1 Review of submodular function theory

Lovász extension. Given any set-function F, one can define its *Lovász extension* $f : \mathbb{R}_+^d \to \mathbb{R}$, as follows: given $w \in \mathbb{R}_+^d$, we can order the components of w in decreasing order $w_{j_1} \geqslant \cdots \geqslant w_{j_p} \geqslant 0$, the value $f(w)$ is then defined as

$$f(w) = \sum_{k=1}^{p-1} (x_{j_k} - x_{j_{k+1}}) F(\{j_1, \ldots, j_k\}) + x_{j_p} F(\{j_1, \ldots, j_p\}) \quad (6)$$

$$= \sum_{k=1}^{p} w_{j_k} [F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})]. \quad (7)$$

The Lovász extension f is always piecewise-linear, and when F is submodular, it is also convex (see, e.g., Fujishige (2005); Bach (2011)). Moreover, for all $\delta \in \{0,1\}^d$, $f(\delta) = F(\mathrm{Supp}(\delta))$ and f is in that sense an extension of F from vectors in $\{0,1\}^d$ (which can be identified with indicator vectors of sets) to all vectors in $\mathbb{R}_+^d$. Moreover, it turns out that minimizing F over subsets, i.e., minimizing f over $\{0,1\}^d$ is equivalent to minimizing f over $[0,1]^d$ (Edmonds, 2003).

Submodular polyhedron and norm We denote by $\mathcal{P}$ the *submodular polyhedron* (Fujishige, 2005), defined as the set of $s \in \mathbb{R}_+^d$ such that for all $A \subset V$, $s(A) \leqslant F(A)$, i.e., $\mathcal{P} = \{s \in \mathbb{R}_+^d, \ \forall A \subset V, \ s(A) \leqslant F(A)\}$, where we use the notation $s(A) = \sum_{k \in A} s_k$. With our previous definitions, the submodular polyhedron is just the canonical polyhedron associated with a submodular function. One important result in submodular analysis is that, if F is a nondecreasing submodular function, then we have a representation of f as a maximum of linear functions (Fujishige, 2005; Bach, 2011), i.e., for all $w \in \mathbb{R}_+^d$,

$$f(w) = \max_{s \in \mathcal{P}} \ w^\top s. \quad (8)$$

We recognize here that the Lovász extension of a submodular function F is directly related to the norm Ω_∞^F in that $f(|w|) = \Omega_\infty^F(w)$ for all $w \in \mathbb{R}^d$.

Greedy algorithm Instead of solving a linear program with $d + 2^d$ constraints, a solution s to (8) may be obtained by the following algorithm (a.k.a. "greedy algorithm"): order the components of w in decreasing order $w_{j_1} \geqslant \cdots \geqslant w_{j_d}$, and then take for all $k \in V$, $s_{j_k} = F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})$. Moreover, if $w \in \mathbb{R}^d$ has some negative components, then, to obtain a solution to $\max_{s \in \mathcal{P}} \ w^\top s$, we can take s_{j_k} to be simply equal to zero for all k such that w_{j_k} is negative (Edmonds, 2003).

Contraction and restriction of a submodular function. Given a submodular function F and a set J, two related functions, which are submodular as well, will play a crucial role both

algorithmically and for the theoretical analysis of the norm. Those are the *restriction* of F to a set J, denoted F_J, and the *contraction* of F on J, denoted F^J. They are defined respectively as

$$F_J : A \mapsto F(A \cap J) \qquad \text{and} \qquad F^J : A \mapsto F(A \cup J) - F(A).$$

Both F_J and F^J are submodular if F is.

In particular the norms $\Omega_p^{F_J} : \mathbb{R}^J \to \mathbb{R}_+$ and $\Omega_p^{F^J} : \mathbb{R}^{J^c} \to \mathbb{R}_+$ associated respectively with F_J and F^J will be useful to "decompose" Ω_p^F in the sequel. We will denote these two norms by Ω_J and Ω^J for short. Note that their domains are not $\mathbb{R}^d$ but the vectors with support in J and J^c respectively.

Stable sets. Another concept which will be key in this section is that of *stable set*. A set A is said *stable* if it cannot be augmented without increasing F, i.e., if for all sets $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$. If F is strictly increasing (such as for the cardinality), then all sets are stable. The set of stable sets is closed by intersection. In the case $p = \infty$, Bach (2011) has shown that these stable sets were the only allowed sparsity patterns.

Separable sets. A set A is separable if we can find a partition of A into $A = B_1 \cup \cdots \cup B_k$ such that $F(A) = F(B_1) + \cdots + F(B_k)$. A set A is inseparable if it is not separable. As shown in Edmonds (2003), the submodular polytope $\mathcal{P}$ has full dimension d as soon as F is strictly positive on all singletons, and its faces are exactly the sets $\{s(A) = F(A)\}$ for stable *and* inseparable sets A. With the terminology that we introduced in Section 2.3, this means that the core set of F is the set $\mathcal{D}_F$ of its stable and inseparable sets. In other words, we have $\mathcal{P} = \{s \in \mathbb{R}^d,\ \forall A \in \mathcal{D}_F, s(A) \leqslant F(A)\}$. The core set will clearly play a role when deriving concentration inequalities in Section 6.5. For the cardinality function, stable and inseparable sets are singletons.

6.2 Submodular function and lower combinatorial envelope

A few comments are in order to confront submodularity to the previously introduced notions associated with cover-sets, and lower and upper combinatorial envelopes. We have showed that $F_-(A) = \Omega_\infty(1_A)$. But for a submodular function $\Omega_\infty(1_A) = f(1_A) = F(A)$ since f is the Lovász extension of F. This shows that a submodular function is its own lower combinatorial envelope. However the converse is not true: a lower combinatorial envelope is not submodular in general. Indeed, in example 2, we have $F_-(\{1,2\}) + F_-(\{2,3\}) \ngeq F_-(\{2\}) + F_-(\{1,2,3\})$.

The core set of a submodular function is the set $\mathcal{D}_F$ of its stable and inseparable sets, which implies that F can be retrieved as the value of the minimal fractional weighted set cover the sets $A \in \mathcal{D}_F$ with weights $F(A)$.

6.3 Optimization algorithms for the submodular case

In the context of sparsity and structured sparsity, *proximal methods* have emerged as methods of choice to design efficient algorithm to minimize objectives of the form $f(w) + \lambda\Omega(w)$, where f is a smooth function with Lipschitz gradients and Ω is a proper convex function (Bach et al., 2012). In a nutshell, their principle is to linearize f at each iteration and to solve the problem

$$\min_{w \in \mathbb{R}^d} \nabla f(w_t)^\top (w - w_t) + \frac{L}{2}\|w - w_t\|^2 + \lambda\Omega(w),$$

for some constant L. This problem is a special case of the so-called *proximal problem*:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - z\|_2^2 + \lambda \Omega_p(w). \tag{9}$$

The function mapping z to the solution of the above problem is called *proximal operator*. If this proximal operator can be computed efficiently, then proximal algorithm provide good rates of convergence especially for strongly convex objectives. We show in this section that the structure of submodular functions can be leveraged to compute efficiently Ω_p, Ω_p^* and the proximal operator.

6.3.1 Computation of Ω_p and Ω_p^*.

A simple approach to compute the norm is to maximize in κ in the variational formulation (6). This can be done efficiently using for example a *conditional gradient* algorithm, given that maximizing a linear form over the submodular polyhedron is done easily with the *greedy algorithm* (see Section 6.1).

We will propose another algorithm to compute the norm based on the so-called *decomposition algorithm*, which is a classical algorithm of the submodular analysis literature that makes it possible to minimize a separable convex function over the submodular polytope efficiently (see, e.g., Bach, 2011, Section 8.6).

Since the dual norm is defined as $\Omega_p^*(s) = \max_{A \subset V, A \neq \varnothing} \frac{\|s_A\|_q}{F(A)^{1/q}}$, to compute it from s, we need to maximize efficiently over A, which can be done, for submodular functions, through a sequence of submodular function minimizations (see, e.g., Bach, 2011, Section 8.4).

6.3.2 Computation of the proximal operator

Using Eq. (6), we can reformulate problem (9) as

$$\begin{aligned}
\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - z\|_2^2 + \lambda \Omega_p(w) &= \min_{w \in \mathbb{R}^d} \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \frac{1}{2}\|w - z\|_2^2 + \lambda \sum_{i \in V} \kappa_i^{1/q} |w_i| \\
&= \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \sum_{i \in V} \min_{w_i \in \mathbb{R}} \left\{ \frac{1}{2}(w_i - z_i)^2 + \lambda \kappa_i^{1/q} |w_i| \right\} \\
&= \max_{\kappa \in \mathbb{R}_+^d \cap \mathcal{P}} \sum_{i \in V} \psi_i(\kappa_i),
\end{aligned}$$

with $\psi_i : \kappa_i \mapsto \min_{w_i \in \mathbb{R}} \left\{ \frac{1}{2}(w_i - z_i)^2 + \lambda \kappa_i^{1/q} |w_i| \right\}$.

Thus, solving the proximal problem is equivalent to maximizing a concave separable function $\sum_i \psi_i(\kappa_i)$ over the submodular polytope. For a submodular function, this can be solved with a "divide and conquer" strategy which takes the form of the so-called *decomposition algorithm* involving a sequence of submodular function minimizations (see Groenevelt, 1991; Bach, 2011). This yields an algorithm which finds a decomposition of the norm and applies recursively the proximal algorithm to the two parts of the decomposition corresponding respectively to a *restriction* and a *contraction* of the submodular function. We explicit this algorithm as Algorithm 1 for the case $p = 2$.

Applying this decomposition algorithm in the special case where $\lambda = 0$ yields a decomposition algorithm, namely Algorithm E.2, to compute the norm itself (see appendix E.2).

Algorithm 1 Computation $x = \text{Prox}_{\lambda\Omega_2^F}(z)$

Require: $z \in \mathbb{R}^d$, $\lambda > 0$
Let $A = \{j \mid z_j \neq 0\}$
if $A \neq V$ **then**
  Set $x_A = \text{Prox}_{\lambda\Omega_2^{F_A}}(z_A)$
  Set $x_{A^c} = 0$
  return x by concatenating x_A and x_{A^c}
end if
Let $t \in \mathbb{R}^d$ with $t_i = \frac{z_i^2}{\|z\|_2} F(V)$
Find A minimizing the submodular function $F - t$
if $A = V$ **then**
  return $x = \big(\|z\|_2 - \lambda\sqrt{F(V)}\big)_+ \frac{z}{\|z\|_2}$
end if
Let $x_A = \text{Prox}_{\lambda\Omega_2^{F_A}}(z_A)$
Let $x_{A^c} = \text{Prox}_{\lambda\Omega_2^{F^A}}(z_{A^c})$
return x by concatenating x_A and x_{A^c}

6.4 Weak and local decomposability of the norm for submodular functions.

The work of Negahban et al. (2010) has shown that when a norm is *decomposable with respect to a pair of subspaces* A and B, meaning that for all $\alpha \in A$ and $\beta \in B^\perp$ we have $\Omega(\alpha+\beta) = \Omega(\alpha)+\Omega(\beta)$, a common proof scheme allows to show support recovery results and fast rates of convergence in prediction error. For the norms we are considering, this type of assumption would be too strong. Instead, we follow the analysis of Bach (2010) which considered the case $p = \infty$ and which only requires some weaker form of decomposability. The decompositions involve Ω_J and Ω^J which are respectively the norms associated with the *restriction* and the *contraction* of the submodular function F to or on the set J.

Concretely, let $c = \frac{\tilde{m}}{M}$ with $M = \max_{k\in V} F(\{k\})$ and

$$\tilde{m} = \min_{A,k} \; F(A \cup \{k\}) - F(A) \;\; \text{s.t.} \;\; F(A \cup \{k\}) > F(A).$$

Then we have:

Proposition 4. *(Weak and local decomposability)*

Weak decomposability. *For any set J and any $w \in \mathbb{R}^d$, we have*

$$\Omega(w) \geq \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

Local decomposability. *Let $K = Supp(w)$ and J the smallest stable set containing K, if $\|w_{J^c}\|_p \leq c^{1/p} \min_{i\in K} |w_i|$, then*

$$\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

Note that when $p = \infty$, if $J = K$, the condition becomes $\min_{i\in J} |w_i| \geqslant \max_{i\in J^c} |w_i|$, and we recover exactly the corresponding result from Bach (2010).

This proposition shows that a sort of reverse triangular inequality involving the norms Ω, Ω_J and Ω^J always holds and that if there is a sufficiently large positive gap between the values of w on J and on its complement then Ω can be written as a separable function on J and J^c.

6.5 Theoretical analysis for submodular functions

In this section, we consider a fixed design matrix $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$ a vector of random responses. Given $\lambda > 0$, we define $\hat{w}$ as a minimizer of the regularized least-squares cost:

$$\min_{w\in\mathbb{R}^d} \tfrac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w). \tag{10}$$

We study the sparsity-inducing properties of solutions of (10), i.e., we determine which patterns are allowed and which sufficient conditions lead to correct estimation.

We assume that the linear model is well-specified and extend results from Zhao and Yu (2006) for sufficient support recovery conditions and from Negahban et al. (2010) for estimation consistency, which were already derived by Bach (2010) for $p = \infty$. The following propositions allow us to retrieve and extend well-known results for the ℓ_1-norm.

Denote by ρ the following constant:

$$\rho = \min_{A\subset B, F(B)>F(A)} \frac{F(B) - F(A)}{F(B\backslash A)} \in (0, 1].$$

The following proposition extends results based on support recovery conditions (Zhao and Yu, 2006):

Proposition 5 (Support recovery). *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$. Denote by J the smallest stable set containing the support $Supp(w^*)$ of w^*. Define $\nu = \min_{j, w_j^*\neq 0} |w_j^*| > 0$ and assume $\kappa = \lambda_{\min}(Q_{JJ}) > 0$.*

*If the following **generalized Irrepresentability Condition** holds:*

$$\exists \eta > 0, \qquad (\Omega^J)^*\Big(\big(\Omega_J(Q_{JJ}^{-1}Q_{Jj})\big)_{j\in J^c}\Big) \leqslant 1 - \eta,$$

then, if $\lambda \leqslant \frac{\kappa\nu}{2|J|^{1/p}F(J)^{1-1/p}}$, the minimizer $\hat{w}$ is unique and has support equal to J, with probability larger than $1 - 3\,\mathbb{P}\big(\Omega^*(z) > \frac{\lambda\eta\rho\sqrt{n}}{2\sigma}\big)$, where z is a multivariate normal with covariance matrix Q.

In terms of prediction error the next proposition extends results based on restricted eigenvalue conditions (see, e.g. Negahban et al., 2010).

Proposition 6 (Consistency). *Assume that $y = Xw^* + \sigma\varepsilon$, where ε is a standard multivariate normal vector. Let $Q = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$. Denote by J the smallest stable set containing the support $Supp(w^*)$ of w^*.*

*If the following Ω_J**-Restricted Eigenvalue condition** holds:*

$$\forall \Delta \in \mathbb{R}^d, \qquad \big(\,\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)\,\big) \quad \Rightarrow \quad \big(\,\Delta^\top Q\Delta \geqslant \kappa\,\Omega_J(\Delta_J)^2\,\big),$$

then we have

$$\Omega(\hat{w} - w^*) \leqslant \frac{24^2\lambda}{\kappa\rho^2} \qquad \text{and} \qquad \frac{1}{n}\|X\hat{w} - Xw^*\|_2^2 \leqslant \frac{36\lambda^2}{\kappa\rho^2},$$

with probability larger than $1 - \mathbb{P}\big(\Omega^*(z) > \frac{\lambda\rho\sqrt{n}}{2\sigma}\big)$ where z is a multivariate normal with covariance matrix Q.

The concentration of the values of $\Omega^*(z)$ for z is a multivariate normal with covariance matrix Q can be controlled via the following result.

Proposition 7. *Let z be a normal variable with covariance matrix Q that has unit diagonal. Let $\mathcal{D}_F$ be the set of stable inseparable sets. Then*

$$\mathbb{P}\bigg(\Omega^*(z) \geqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}} + u\max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}\bigg) \leqslant e^{-u^2/2}. \tag{11}$$

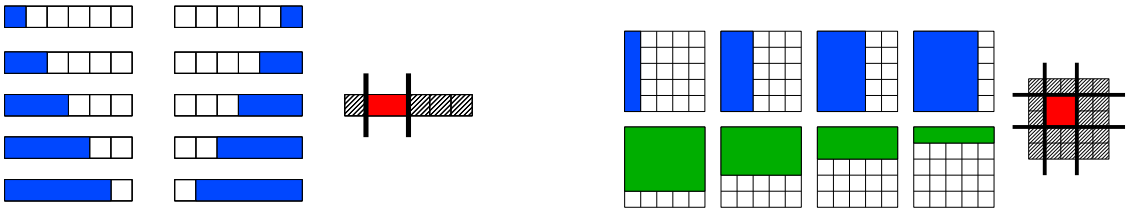

Figure 4: Set $\mathcal{G}$ of overlapping groups defining the norm proposed by Jenatton et al. (2011a) (set in blue or green and their complements) and an example of corresponding induced sparsity patterns (in red), respectively for interval patterns in 1D (left) and for rectangular patterns in 2D (right).

7 Experiments

7.1 Setting

To illustrate the results presented in this paper we consider the problem of estimating the support of a parameter vector $w \in \mathbb{R}^d$, when its support is assumed either

(i) to form an *interval* in $[\![1, d]\!]$ or

(ii) to form a *rectangle* $[\![k_{\min}, k_{\max}]\!] \times [\![k'_{\min}, k'_{\max}]\!] \subset [\![1, d_1]\!] \times [\![1, d_2]\!]$, with $d = d_1 d_2$.

These two settings were considered in Jenatton et al. (2011a). These authors showed that, for both types of supports, it was possible to construct an ℓ_1/ℓ_2-norm with overlap based on a well-chosen collection of overlapping groups, so that the obtained estimators almost surely have a support of the correct form. Specifically, it was shown in Jenatton et al. (2011a) that norms of the form $w \mapsto \sum_{B \in \mathcal{G}} \|w_B\|_2$ induce sparsity patterns that are exactly intervals of $V = \{1, \dots, p\}$ if

$$\mathcal{G} = \big\{[1, k] \mid 1 \leq k \leq p\big\} \cup \big\{[k, p] \mid 1 \leq k \leq p\big\},$$

and induce rectangular supports on $V = V_1 \times V_2$ with $V_1 := \{1, \dots, p_1\}$ and $V_2 := \{1, \dots, p_2\}$ if

$$\begin{aligned} \mathcal{G} \quad = \quad & \big\{[\![1, k]\!] \times V_2 \mid 1 \leq k \leq p_1\big\} \cup \big\{[\![k, p_1]\!] \times V_2 \mid 1 \leq k \leq p_1\big\} \\ & \cup \big\{V_1 \times [\![1, k]\!] \mid 1 \leq k \leq p_2\big\} \cup \big\{V_1 \times [\![k, p_2]\!]\big\} \mid 1 \leq k \leq p_2\big\}. \end{aligned}$$

These sets of groups are illustrated on Figure 4, and, for the first case, the set $\mathcal{G}$ has already discussed in Example 3 to define a modified range function which is submodular.

Moreover, the authors showed that with a weighting scheme leading to a norm of the form $w \mapsto \sum_{B \in \mathcal{G}} \|w_B \circ d^B\|$, where $\circ$ denotes the Hadamard product and $d^B \in \mathbb{R}^d_+$ is a certain vector of weights designed specifically for these case[7] it is possible to obtain compelling empirical results in terms of support recovery, especially in the 1D case.

Interval supports. From the point of view of our work, that is, approaching the problem in terms of combinatorial functions, for supports constrained to be intervals, it is natural to consider the range function as a possible form of penalty: $F_0(A) := \text{range}(A) = i_{\max}(A) - i_{\min}(A) + 1$. Indeed the range function assigns the same penalty to sets with the same range, regardless of whether these sets are connected or have "holes"; this clearly favors intervals since they are exactly the sets with

[7] We refer the reader to the paper for the details.

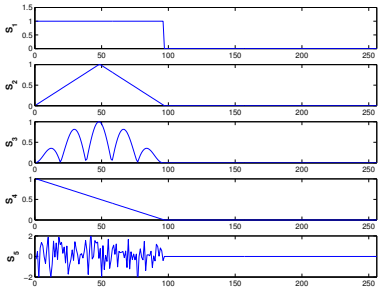

Figure 5: Examples of the shape of the signals used to define the amplitude of the coefficients of w on the support. Each plot represents the value of w_i as a function of i. The first (w constant on the support), third ($w_i = g(ci)$ with $g : x \mapsto |\sin(x)\sin(5x)|$) and last signal ($w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$) are the ones used in reported results.

the largest support for a given value of the penalty. Unfortunately, as discussed in the Example 1 of Section 2.2, the combinatorial lower envelope of the range function is $A \mapsto |A|$, the cardinality function, which implies that $\Omega_p^{F_0}$ is just the ℓ_1-norm: in this case, the structure implicitly encoded in F_0 is lost through the convex relaxation.

However, as mentioned by Bach (2010) and discussed in Example 3 the function F_r defined by $F_r(A) = d - 1 + \text{range}(A)$ for $A \neq \varnothing$ and $F(\varnothing) = 0$ is submodular, which means that $\Omega_p^{F_r}$ is a tight relaxation and that regularizing with it leads to tractable convex optimization problems.

Rectangular supports. For the case of rectangles on the grid, a good candidate is the function F_2 with $F_2(A) = F_r(\Pi_1(A)) + F_r(\Pi_2(A))$ with $\Pi_i(A)$ the projection of the set A along the ith axis of the grid.

This makes of $\Omega_p^{F_r}$ and $\Omega_p^{F_2}$ two good candidates to estimate a vector w whose support matches respectively the two described a priori.

7.2 Methodology

We consider a simple regression setting in which $w \in \mathbb{R}^d$ is a vector such that $\text{Supp}(w)$ is either an interval on $[1, d]$ or a rectangle on a fixed $2D$ grid. We draw the design matrix $X \in \mathbb{R}^{n \times d}$ and a noise vector $\epsilon \in \mathbb{R}^n$ both with i.i.d. standard Gaussian entries and compute $y = Xw + \epsilon$. We then solve problem (10), with Ω chosen in turn to be the ℓ_1-norm (Lasso), the elastic net, the norms Ω_p^F for $p \in \{2, \infty\}$ and F chosen to be F_r or F_2 in $1D$ and $2D$ respectively; we consider also the overlapping ℓ_1/ℓ_2-norm proposed by Jenatton et al. (2011a) and the weighted overlapping ℓ_1/ℓ_2-norm proposed by the same authors, i.e., $\Omega(w) = \sum_{B \in \mathcal{G}} \|w_B \circ d^B\|_2$ with the same notations as before[8].

We assess the estimators obtained through the different regularizers both in terms of support recovery and in terms of mean-squared error in the following way: assuming that held out data permits to choose an optimal point on the regularization path obtained with each norm, we determine along each such path, the solution which either has a support with minimal Hamming distance to the true support or the solution which as the best ℓ_2 distance, and we report the corresponding distances as a function the sample size on Figures 6 and 7 respectively for the 1D and the 2D case.

[8]Note that we do not need to compare with an $\ell_i nfty$ counterpart of the unweighted norm considered in Jenatton et al. (2011a) since for $p = \infty$ the unweighted ℓ_1/ℓ_∞ norm defined with the same collection $\mathcal{G}$ is exactly the norm $\Omega_\infty^{F_r}$: this follows from the form of F_r as defined in Example 3 and the preceding discussion.

Finally, we assess the incidence of the fluctuation in amplitude of the coefficients in the vector w generating the data: we consider different cases among which:

(i) the case where w has a constant value on the support,

(ii) the case where w_i varies as a modulated cosine, with $w_i = g(ci)$ for c a constant scaling and $g : x \mapsto |\cos(x)\cos(5x)|$

(iii) the case where w_i is drawn i.i.d. from a standard normal distribution.

These cases (and two others for which we do not report results) are illustrated on Figure 5.

7.3 Results

Results reported for the Hamming distances in the left columns of Figures 6 and 7 show that the norms $\Omega_2^{F_r}$ and $\Omega_2^{F_2}$ perform quite well for support recovery overall and tend to outperform significantly their ℓ_∞ counterpart in most cases. In 1D, several norms achieve reasonably small Hamming distance, including the ℓ_1-norm, the norm $\Omega_2^{F_r}$ and the weighted overlapping ℓ_1/ℓ_2-norm although the latter clearly dominates for small values of n.

In 2D, $\Omega_2^{F_2}$ leads clearly to smaller Hamming distances than other norms for the larger values of n, while is outperformed by the ℓ_1-norm for small sample sizes. It should be noted that neither $\Omega_\infty^{F_2}$ nor the weighted overlapping ℓ_1/ℓ_2-norm that performed so well in 1D achieve good results.

The performance of the ℓ_2 relaxation tends to be comparatively better when the vector of parameter w has entries that vary a lot, especially when compared to the ℓ_∞ relaxation. Indeed, the choice of the value of p for the relaxation can be interpreted as encoding a prior on the joint distribution of the amplitudes of the w_i: as discussed before, and as illustrated in Bach (2010) the unit balls for the ℓ_∞ relaxations display additional "edges and corners" that lead to estimates with clustered values of $|w_i|$, corresponding to an priori that many entries in w have identical amplitudes. More generally, large values of p correspond to the prior that the amplitude varies little while their vary more significantly for small p.

The effect of this other type of a priori encoded in the regularization is visible when considering the performance in terms of ℓ_2 error. Overall, both in 1D and 2D all methods perform similarly in ℓ_2 error, except that when w is constant on the support, the ℓ_∞ relaxations $\Omega_\infty^{F_r}$ and $\Omega_\infty^{F_2}$ perform significantly better, and this is the case most likely because the additional "corners" of these norms induce some pooling of the estimates of the value of the w_i, which improves their estimation. By contrast it can be noted that when w is far from constant the ℓ_∞ relaxations tend to have slightly larger least-square errors, while, on contrary, the ℓ_1-regularisation tends to be among the better performing methods.

8 Conclusion

We proposed a family of convex norms defined as relaxations of penalizations that combine a combinatorial set-function with an ℓ_p-norm. Our formulation allows to recover in a principled way classical sparsity inducing regularizations such as ℓ_1, ℓ_1/ℓ_p-norms or ℓ_p/ℓ_1-norms. In addition, it establishes the the latent group Lasso is the tightest relaxation of block-coding penalties.

There are several directions for future research. First, it would be of interest to determine for which combinatorial functions beyond submodular ones, efficient algorithms and consistency results can

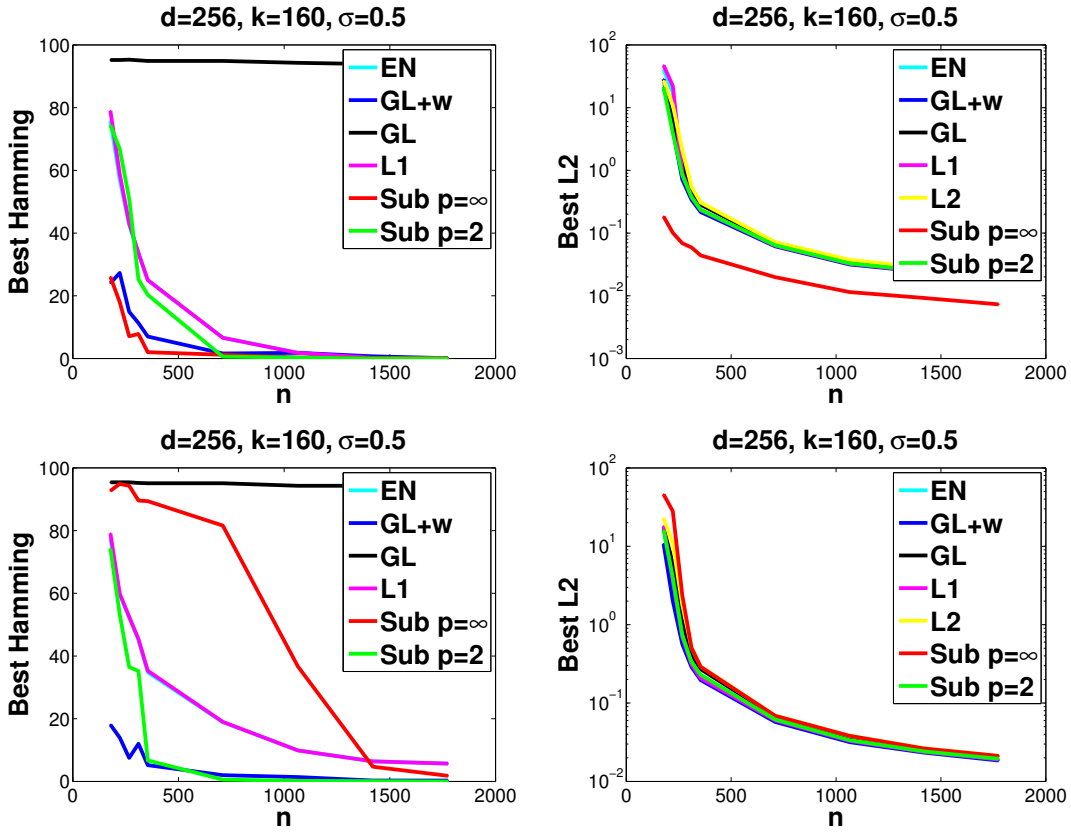

Figure 6: Best Hamming distance (left column) and best least square error (right column) to the true parameter vector w^*, among all vectors along the regularization path of a least square regression regularized with a given norm, for different patterns of values of w^*. The different regularizers compared include the Lasso (L1), Ridge (L2), the elastic net (EN), the unweighted (GL) and weighted (GL+w) ℓ_1/ℓ_2 regularizations proposed by Jenatton et al. (2011a), the norms Ω_2^F (Sub $p = 2$) and Ω_∞^F (Sub $p = \infty$) for a specified function F. (first row) Constant signal supported on an interval, with an a priori encoded by the combinatorial function $F : A \mapsto d - 1 + \text{range}(A)$. (second row) Same setting with a signal w^* supported by an interval consisting of coefficients w_i^* drawn from a standard Gaussian distribution. In each case, the dimension is $d = 256$, the size of the true support is $k = 160$, the noise level is $\sigma = 0.5$ and signal amplitude $\|w\|_\infty = 1$.

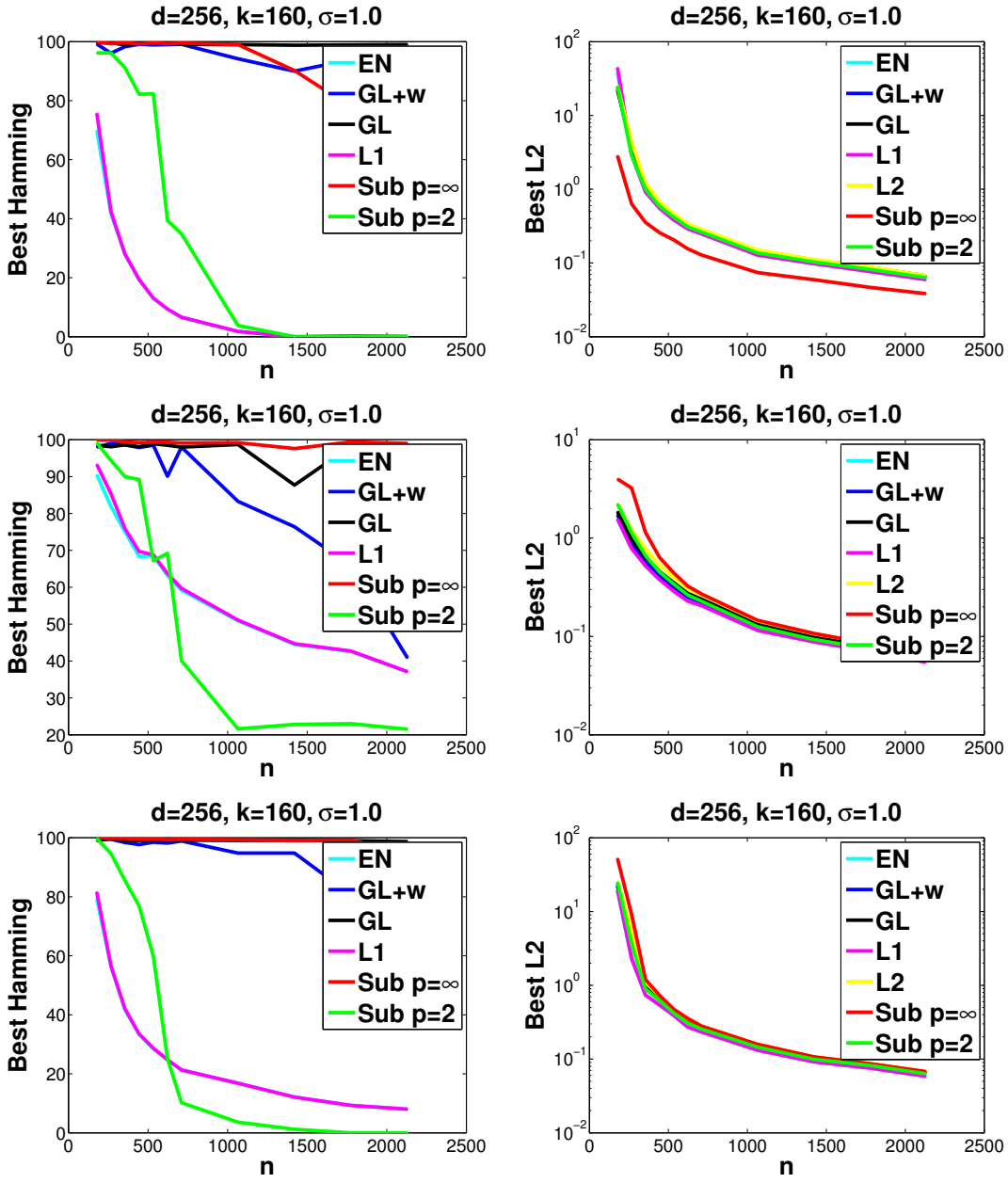

Figure 7: Best Hamming distance (left column) and best least square error (right column) to the true parameter vector w^*, among all vectors along the regularization path of a least square regression regularized with a given norm, for different patterns of values of w^*. The regularizations compared include the Lasso (L1), Ridge (L2), the elastic net (EN), the unweighted (GL) and weighted (GL+w) ℓ_1/ℓ_2 regularizations proposed by Jenatton et al. (2011a), the norms Ω_2^F (Sub $p = 2$) and Ω_∞^F (Sub $p = \infty$) for a specified function F. Parameter vectors w^* considered here have coefficients that are supported by a rectangle on a grid with size $d_1 \times d_2$ with $d = d_1 d_2$. (first row) Constant signal supported on a rectangle with an a priori encoded by the combinatorial function $F : A \mapsto d_1 + d_2 - 4 + \text{range}(\Pi_1(A)) + \text{range}(\Pi_2(A))$. (second row) Same setting with coefficients of w on the support given as $w^*_{i_1 i_2} = g(c\, i_1) g(c\, i_2)$ for c a positive constant and $g : x \mapsto |\cos(x)\cos(5x)|$. (third row) Same setting with coefficients $w^*_{i_1 i_2}$ drawn from a standard Gaussian distribution. In each case, the dimension is $d = 256$, the size of the true support is $k = 160$, the noise level is $\sigma = 1$ and signal amplitude $\|w\|_\infty = 1$.

be established. Then a sharper analysis of the relative performance of the estimators using different levels of a priori would be needed to answer question such as: When is using a structured a priori likely to yield better estimators? When could it degrade the performance? What is the relation to the performance of an oracle given a specified structured a priori?

Acknowledgements

The authors acknowledge funding from the European Research Council grant SIERRA, project 239993, and would like to thank Rodolphe Jenatton and Julien Mairal for stimulating discussions.

References

F. Bach. Learning with submodular functions: A convex optimization perspective. *Arxiv preprint arXiv:1111.6453*, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundation and Trends in Machine Learning*, 1(4):1–106, 2012.

Francis Bach. Structured sparsity-inducing norms through submodular functions. In *Adv. NIPS*. 2010.

R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory,*, 56(4):1982–2001, 2010.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Arxiv preprint arXiv:1012.0621*, 2010.

J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer, 2003.

S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *Eur. J Oper. Res.*, 54(2):227–236, 1991.

L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497, 2009.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *The JMLR*, 12:3371–3412, 2011.

L. Jacob, G. Obozinski, and J.P. Vert. Group lasso with overlap and graph lasso. In *ICML*, 2009.

R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *JMLR*, 12:2777–2824, 2011a.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *JMLR*, 12:2297–2334, 2011b.

S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *Proc. ICML*, 2010.

L. Lovász. On the ratio of optimal integral and fractional covers. *Discr. Math.*, 13(4):383–390, 1975.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. Technical Report 1204.4539, arXiv, 2012.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *JMLR*, 12:2681–2720, 2011.

C.A. Micchelli, J.M. Morales, and M. Pontil. Regularizers for structured sparsity. *Arxiv preprint arXiv:1010.0556*, 2011. To appear in Advances in Computational Mathematics.

S. Negahban, P. Ravikumar, M.J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Arxiv preprint arXiv:1010.2731*, 2010.

G. Obozinski, L. Jacob, and J.-P. Vert. Group Lasso with overlaps: the Latent Group Lasso approach. *preprint HAL - inria-00628498*, 2011.

J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

R.T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of Lasso. *JMLR*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Ann. of Stat.*, 37(6A):3468–3497, 2009.

Y. Zhou, R. Jin, and S. C. Hoi. Exclusive lasso for multi-task feature selection. In *AISTATS*, 2010.

A Form of primal norm

We provide here a proof of lemma 6 which we first recall:

Lemma (6)**.** *Ω_p and Ω_p^* are dual to each other.*

Proof. Let ω_p^A be the function[9] defined by $\omega_p^A(w) = F(A)^{1/q} \|w_A\|_p \, \iota_{\{v|Supp(v)\subset A\}}(w)$ with ι_B the indicator function taking the value 0 on B and ∞ on B^c. Let K_p^A be the set $K_p^A = \{s \mid \|s_A\|_q^q \leq F(A)\}$. By construction, ω_p^A is the *support function* of K_p^A (see Rockafellar, 1970, sec.13), i.e. $\omega_p^A(w) = \max_{s\in K_p^A} w^\top s$. By construction we have $\{s \mid \Omega_p^*(s) \leq 1\} = \cap_{A\subset V} K_p^A$. But this implies that $\iota_{\{s|\Omega_p^*(s)\leq 1\}} = \sum_{A\subset V} \iota_{K_p^A}$. Finally, by definition of Fenchel-Legendre duality,

$$\Omega_p(w) = \max_{w\in\mathbb{R}^d} w^\top s - \sum_{A\subset V} \iota_{K_p^A}(s),$$

or in words Ω_p is the Fenchel-Legendre dual to the sum of the indicator functions $\iota_{K_p^A}$. But since the Fenchel-Legendre dual of a sum of functions is the *infimal convolution* of the duals of these

[9] Or gauge function to be more precise.

functions (see Rockafellar, 1970, Thm. 16.4 and Corr. 16.4.1, pp. 145-146), and since by definition of a support function $\left(\iota_{K_p^A}\right)^* = \omega_p^A$, then Ω_p is the *infimal convolution* of the functions ω_p^A, i.e.

$$\Omega_p(w) = \inf_{(v^A \in \mathbb{R}^d)_{A \subset V}} \sum_{A \subset V} \omega_p^A(v^A) \quad \text{s.t.} \quad w = \sum_{A \subset V} v^A,$$

which is equivalent to formulation (3). See Obozinski et al. (2011) for a more elementary proof of this result. □

B Example of the Exclusive Lasso

We showed in Section 4.2 that the ℓ_p exclusive Lasso norm, also called ℓ_p/ℓ_1-norm, defined by the mapping $w \mapsto \left(\sum_{G \in \mathcal{G}} \|w_G\|_1^p\right)^{1/p}$, for some partition $\mathcal{G}$, is a norm Ω_p^F providing the ℓ_p tightest convex p.h. relaxation in the sense defined in this paper of a certain combinatorial function F. A computation of the lower combinatorial envelope of that function F yields the function $F_- : A \mapsto \max_{G \in \mathcal{G}} |A \cap G|$.

This last function is also a natural combinatorial function to consider and by the properties of a LCE it has the same convex relaxation. It should be noted that it is however less obvious to show directly that $\Omega_p^{F_-}$ is the ℓ_p/ℓ_1 norm...

We thus show a direct proof of that result since it illustrates how the results on LCE and UCE can be used to analyze norms and derive such results.

Lemma 9. *Let $\mathcal{G} = \{G_1, \dots, G_k\}$ be a partition of V. For $F : A \mapsto \max_{G \in \mathcal{G}} |A \cap G|$, we have $\Omega_\infty^F(w) = \max_{G \in \mathcal{G}} \|w_G\|_1$.*

Proof. Consider the function $f : w \mapsto \max_{G \in \mathcal{G}} \|w_G\|_1$ and the set function $F_0 : A \mapsto f(1_A)$. We have $F_0(A) = \max_{G \in \mathcal{G}} \|1_{A \cap G}\|_1 = F(A)$. But by Lemma 7, this implies that $f(w) \leq \Omega_\infty^F(w)$ since $f = f(|\cdot|)$ is convex positively homogeneous and coordinatewise non-decreasing on $\mathbb{R}_+^d$. We could remark first that since $F(A) = f(1_A) \leq \Omega_\infty^F(1_A) \leq F(A)$, this shows that $F = F_-$ is a lower combinatorial envelope. Now note that

$$(\Omega_\infty^F)^*(s) = \max_{A \subset V, A \neq \varnothing} \min_{G \in \mathcal{G}} \frac{\|s_A\|_1}{|A \cap G|} \geq \max_{A \subset V,\, |A \cap G| = 1,\, G \in \mathcal{G}} \|s_A\|_1 = \sum_{G \in \mathcal{G}} \max_{i \in G} |s_i| = \sum_{G \in \mathcal{G}} \|s_G\|_\infty.$$

This shows that $(\Omega_\infty^F)^*(s) \geq \sum_{G \in \mathcal{G}} \|s_G\|_\infty$, which implies for dual norms that $\Omega_\infty^F(w) \leq f(w)$. Finally, since we showed above the opposite inequality $\Omega_\infty^F = f$ which shows the result. □

C Properties of the norm Ω_p^F when F is submodular

In this section, we first derive upper bounds and lower bounds for our norms, as well as a local formulation as a sum of ℓ_p-norms on subsets of indices.

C.1 Some important inequalities.

We now derive inequalities which will be useful later in the theoretical analysis. By definition, the dual norm satisfies the following inequalities:

$$\frac{\|s\|_\infty}{M^{\frac{1}{q}}} \leq \max_{k\in V} \frac{\|s_{\{k\}}\|_q}{F(\{k\})^{\frac{1}{q}}} \leq \Omega_p^*(s) = \max_{A\subset V, A\neq\varnothing} \frac{\|s_A\|_q}{F(A)^{\frac{1}{q}}} \leq \frac{\|s\|_q}{\min_{A\subset V, A\neq\varnothing} F(A)^{\frac{1}{q}}} \leq \frac{\|s\|_q}{m^{\frac{1}{q}}}, \qquad (12)$$

for $m = \min_{k\in V} F(\{k\})$ and $M = \max_{k\in V} F(\{k\})$. These inequalities imply immediately inequalities for Ω_p (and therefore for f since for $\eta \in \mathbb{R}_+^d$, $f(\eta) = \Omega_\infty(\eta)$):

$$m^{1/q}\|w\|_p \leqslant \Omega_p(w) \leqslant M^{1/q}\|w\|_1.$$

We also have $\Omega_p(w) \leqslant F(V)^{1/q}\|w\|_p$, using the following lower bound for the dual norm: $\Omega_p^*(s) \geqslant \frac{\|s\|_p}{F(V)^{1/q}}$.

Since by submodularity, we in fact have $M = \max_{A,k\notin A} F(A\cup\{k\}) - F(A)$, it makes sense to introduce $\tilde{m} = \min_{A,k,F(A\cup\{k\})>F(A)} F(A\cup\{k\}) - F(A) \leq m$. Indeed, we consider in Section 6.5 the norm $\Omega_{p,J}$ (resp. Ω_p^J) associated with *restrictions* of F to J (resp. *contractions* of F on J) and it follows from the previous inequalities that for all $J \subset V$, we have:

$$\tilde{m}^{1/q}\|w\|_p \leqslant m^{1/q}\|w\|_p \leqslant \Omega_{p,J}(w) \leqslant M^{1/q}\|w\|_1 \qquad \text{and} \qquad \tilde{m}^{1/q}\|w\|_p \leqslant \Omega_p^J(w) \leqslant M^{1/q}\|w\|_1.$$

C.2 Some optimality conditions for η.

While exact necessary and sufficient conditions for η to be a solution of Eq. (6) would be tedious to formulate precisely, we provide three necessary and two sufficient conditions, which together characterize a non-trivial subset of the solutions, which will be useful in the subsequent analysis.

Proposition 8 (Optimality conditions for η)**.** *Let F be a non-increasing submodular function. Let $p > 1$ and $w \in \mathbb{R}^d$, $K = Supp(w)$ and J the smallest stable set containing K. Let $H(w)$ the set of minimizers of Eq. (6). Then,*

(a) the set $\{\eta_K,\ \eta \in H(w)\}$ is a singleton with strictly positive components, which we denote $\{\eta_K(w)\}$, i.e., Eq. (6) uniquely determines η_K.

(b) For all $\eta \in H(w)$, then $\eta_{J^c} = 0$.

(c) If $A_1 \cup \cdots \cup A_m$ are the ordered level sets of η_K, i.e., η is constant on each A_j and the values on A_j form a strictly decreasing sequence, then $F(A_1 \cup \cdots \cup A_j) - F(A_1 \cup \cdots \cup A_{j-1}) > 0$ and the value on A_j is equal to $\eta^{A_j}(w) = \frac{\|w_{A_j}\|_p}{[F(A_1\cup\cdots\cup A_j) - F(A_1\cup\cdots\cup A_{j-1})]^{1/p}}$.

(d) If η_K is equal to $\eta_K(w)$, $\max_{k\in J\setminus K} \eta_k \leqslant \min_{k\in K} \eta_k(w)$, and $\eta_{J^c} = 0$, then $\eta \in H(w)$.

(e) There exists $\eta \in H(w)$ such that $\frac{\min_{i\in K}|w_i|}{M^{1/p}} \leqslant \min_{j\in J} \eta_j \leqslant \max_{j\in J} \eta_j \leqslant \frac{\|w\|_p}{m^{1/p}}$.

Proof. (a) Since f is non-decreasing with respect to each of its argument, for any $\eta \in H(w)$, we have $\eta' \in H(w)$ for η' defined through $\eta'_K = \eta_K$ and $\eta_{K^c} = 0$. The set of values of η_K for $\eta \in H(w)$ is therefore the set of solutions problem (6) restricted to K. The latter problem has a unique solution as a consequence of the strict convexity on $\mathbb{R}_+^*$ of $\eta_j \mapsto \frac{|w_j|^p}{\eta_j^{p-1}}$.

(b) If there is $j \in J^c$ such that $\eta \in H(w)$ and $\eta_j \neq 0$, then (since $w_j = 0$) because f is non-decreasing with respect to each of its arguments, we may take η_j infinitesimally small and all other

η_k for $k \in K^c$ equal to zero, and we have $f(\eta) = f_K(\eta_K(w)) + \eta_j[F(K \cup \{j\}) - F(K)]$. Since $F(K \cup \{j\}) - F(K) \geqslant F(J \cup \{j\}) - F(J) > 0$ (because J is stable), we have $f(\eta) > f_K(\eta_K(w))$, which is a contradiction.

(c) Given the ordered level sets, we have $f(\eta) = \sum_{j=1}^m \eta^{A_j}[F(A_1 \cup \cdots \cup A_j) - F(A_1 \cup \cdots \cup A_{j-1})]$, which leads to a closed-form expression $\eta^{A_j}(w) = \frac{\|w_{A_j}\|_p}{[F(A_1 \cup \cdots \cup A_j) - F(A_1 \cup \cdots \cup A_{j-1})]^{1/p}}$. If $F(A_1 \cup \cdots \cup A_j) - F(A_1 \cup \cdots \cup A_{j-1}) = 0$, since $\|w_{A_j}\|_p > 0$, we have η^{A_j} as large as possible, i.e., it has to be equal to $\eta^{A_{j-1}}$, thus it is not a possible ordered partition.

(d) With our particular choice for η, we have $\sum_{i \in V} \frac{1}{p} \frac{|w_i|^p}{\eta_i^{p-1}} + \frac{1}{q} f(\eta) = \Omega_K(w_K)$. Since we always have $\Omega(w) \geqslant \Omega_K(w_K)$, then η is optimal in Eq. (6).

(e) We take the largest elements from (d) and bounds the components of η_K using (c). □

Note that from property (c), we can explicit the value of the norm as:

$$\Omega_p(w) = \sum_{j=1}^k (F(A_1 \cup \ldots \cup A_j) - F(A_1 \cup \ldots \cup A_{j-1}))^{\frac{1}{q}} \|w_{A_j \setminus A_{j-1}}\|_p \tag{13}$$

$$= \Omega_{p,A_1}(w_{A_1}) + \sum_{j=2}^k \Omega_{p,A_j}^{A_{j-1}}(w_{A_j \setminus A_{j-1}}) \tag{14}$$

where $\Omega_{p,B}^A$ is the norm associated with the contraction on A of F restricted to B.

D Proof of Proposition 4 (Decomposability)

Concretely, let $c = \frac{\tilde{m}}{M}$ with $M = \max_{k \in V} F(\{k\})$ and

$$\tilde{m} = \min_{A,k} \; F(A \cup \{k\}) - F(A) \;\text{ s.t. }\; F(A \cup \{k\}) > F(A)$$

Proposition (4. Weak and local Decomposability)**.** *(a) For any set J and any $w \in \mathbb{R}^d$, we have*

$$\Omega(w) \geq \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

(b) Assume that J is stable, and $\|w_{J^c}\|_p \leq c^{1/p} \min_{i \in J} |w_i|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$.
(c) Assume that K is non stable and J is the smallest stable set containing K, and that $\|w_{J^c}\|_p \leq c^{1/p} \min_{i \in K} |w_i|$, then $\Omega(w) = \Omega_J(w_J) + \Omega^J(w_{J^c})$.

Proof. We first prove the first statement (a): If $\|s_{A \cap J}\|_p^p \leq F(A \cap J)$ and $\|s_{A \cap J^c}\|_p^p \leq F(A \cup J) - F(J)$ then by submodularity we have $\|s_A\|_p^p \leq F(A \cap J) + F(A \cup J) - F(J) \leq F(A)$. The submodular polyhedra associated with F_J and F^J are respectively defined by

$$P(F_J) = \{s \in \mathbb{R}^d, \, \text{Supp}(s) \subset J, \, s(A) \leq F(A), \, A \subset J\} \quad \text{and}$$
$$P(F^J) = \{s \in \mathbb{R}^d, \, \text{Supp}(s) \subset J^c, \, s(A) \leq F(A \cup J) - F(J)\}$$

Denoting $s^{\circ p} := (s_1^p, \ldots, s_d^p)$, we therefore have

$$\Omega(w) = \max_{\{|s^{\circ p}| \in P(F)\}} s^\top w \geq \max_{\{|s_J^{\circ p}| \in P(F_J), \, |s_{J^c}^{\circ p}| \in P(F^J)\}} s^\top w = \Omega_J(w_J) + \Omega^J(w_{J^c}).$$

In order to prove (b), we consider an optimal η_J for w_J and Ω_J and an optimal η_{J^c} for Ω^J. Because of our inequalities, and because we have assume that J is stable (so that the value m for Ω^J is indeed lower bounded by $\tilde{m}$), we have $\|\eta_{J^c}\|_\infty \leqslant \frac{\|w_{J^c}\|_p}{\tilde{m}^{1/p}}$. Moreover, we have $\min_{j\in J}\eta_j \geqslant \frac{\min_{i\in J}|w_i|}{M^{1/p}}$ (inequality proved in the main paper). Thus when concatenating η_J and η_{J^c} we obtain an optimal η for w (since then the Lovász extension decomposes as a sum of two terms), hence the desired result.

In order to prove (c), we simply notice that since $F(J) = F(K)$, the value of $\eta_{J\setminus K}$ is irrelevant (the variational formulation does not depend on it), and we may take it equal to the largest known possible value, i.e., one which is largest than $\frac{\min_{i\in J}|w_i|}{M^{1/p}}$, and the same reasoning than for (b) applies. □

Note that when $p = \infty$, the condition in (b) becomes $\min_{i\in J}|w_i| \geqslant \max_{i\in J^c}|w_i|$, and we recover exactly the corresponding result from Bach (2010).

E Algorithmic results

E.1 Proof of Algorithm 1

Algorithm 1 is a particular instance of the decomposition algorithm for the optimization of a convex function over the submodular polyhedron (see e.g. section 6.1 of Bach (2011)) Indeed denoting $\psi_i(\kappa_i) = \min_{w_i\in\mathbb{R}} \frac{1}{2}(w_i - z_i)^2 + \lambda\kappa_i^{1/q}|w_i|$, the computation of the proximal operator amounts to solving in κ the problem

$$\max_{\kappa\in\mathbb{R}_+^d\cap\mathcal{P}} \sum_{i\in V}\psi_i(\kappa_i)$$

Following the decomposition algorithm, one has to solve first

$$\begin{aligned}
&\max_{\kappa\in\mathbb{R}_+^d} \sum_{i\in V}\psi_i(\kappa_i) \quad \text{s.t.} \quad \sum_{i\in V}\kappa_i = F(V)\\
=\ &\min_{w\in\mathbb{R}^d}\max_{\kappa\in\mathbb{R}_+^d} \frac{1}{2}\|w-z\|_2^2 + \sum_{i\in V}\kappa_i^{1/q}|w_i| \quad \text{s.t.} \quad \sum_{i\in V}\kappa_i = F(V)\\
=\ &\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w-z\|_2^2 + \lambda F(V)^{1/q}\|w\|_p,
\end{aligned}$$

where the last equation is obtained by solving the maximization problem in κ, which has the unique solution $\kappa_i = F(V)\frac{|w_i|^p}{\|w\|_p^p}$ if $w \neq 0$ and the simplex of solutions $\{\kappa\in\mathbb{R}_+^d \mid \kappa(V) = F(V)\}$ for $w = 0$.

This is solved in closed form for $p = 2$ with $w^* = (\|z\|_2 - \lambda\sqrt{F(V)})_+\frac{z}{\|z\|_2}$ if $z \neq 0$ and $w^* = 0$ else.

In particular since $w^* \propto z$, then $\kappa_i = F(V)\frac{z_i^2}{\|z\|_2^2}$ is always a solution. Following the decomposition algorithm, one then has to find the minimizer of the submodular function $A \mapsto F(A) - \kappa(A)$. Then one needs to solve

$$\min_{\kappa_A\in\mathbb{R}_+^{|A|}\cap\mathcal{P}(F_A)} \sum_{i\in A}\psi_i(\kappa_i) \qquad \text{and} \qquad \min_{\kappa_{V\setminus A}\in\mathbb{R}_+^{|V\setminus A|}\cap\mathcal{P}(F^A)} \sum_{i\in V\setminus A}\psi_i(\kappa_i).$$

Using the expression of ψ_i and exchanging as above the minimization in w and the maximization in κ, one obtains directly that these two problems correspond respectively to the computation of the proximal operators of Ω^{F_A} on z_A and of the proximal operator of Ω^{F^A} on $z_{V\setminus A}$.

The decomposition algorithm is proved to be correct in section 6.1 of Bach (2011) under the assumption that $\kappa_i \mapsto \psi(\kappa_i)$ is a strictly convex function. The functions we consider here are not strongly

convex, and in particular, as mentioned above the solution in κ is not unique in case $w^* = 0$. The proof of Bach (2011) however goes through using any solution of the maximization problem in κ.

E.2 Decomposition algorithm to compute the norm

Applying Algorithm 1 in the special case where $\lambda = 0$ yields a decomposition algorithm to compute the norm itself (see Algorithm E.2).

Algorithm 2 Computation of $\Omega_p^F(z)$

Require: $z \in \mathbb{R}^d$.
Let $A = \{j \mid z_j \neq 0\}$.
if $A \neq V$ **then**
  return $\Omega_p^{F_A}(z_A)$
end if
Let $t \in \mathbb{R}^d$ with $t_i = \frac{|z_i|^p}{\|z\|_p^p} F(V)$
Find A minimizing the submodular function $F - t$
if $A = V$ **then**
  return $F(V)^{1/q}\|x\|_p$
else
  return $\Omega_p^{F_A}(z_A) + \Omega_p^{F^A}(z_{A^c})$
end if

F Theoretical Results

In this section, we prove the propositions on consistency, support recovery and the concentration result of Section 6.5. As there, we consider a fixed design matrix $X \in \mathbb{R}^{n \times p}$ and $y \in \mathbb{R}^n$ a vector of random responses. Given $\lambda > 0$, we define $\hat{w}$ as a minimizer of the regularized least-squares cost:

$$\min_{w \in \mathbb{R}^d} \tfrac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w). \tag{15}$$

F.1 Proof of Proposition 5 (Support recovery)

Proof. We follow the proof of the case $p = \infty$ from Bach (2010). Let $r = \frac{1}{n}X^\top\varepsilon \in \mathbb{R}^d$, which is normal with mean zero and covariance matrix $\sigma^2 Q/n$. We have for any $w \in \mathbb{R}^p$,

$$\Omega(w) \geqslant \Omega_J(w_J) + \Omega^J(w_{J^c}) \geqslant \Omega_J(w_J) + \rho\,\Omega_{J^c}(w_{J^c}) \geqslant \rho\,\Omega(w).$$

This implies that $\Omega^*(r) \geqslant \rho\max\{\Omega_J^*(r_J), (\Omega^J)^*(r_{J^c})\}$.

Moreover, $r_{J^c} - Q_{J^cJ}Q_{JJ}^{-1}r_J$ is normal with covariance matrix

$$\frac{\sigma^2}{n}(Q_{J^cJ^c} - Q_{J^cJ}Q_{JJ}^{-1}Q_{JJ^c}) \preccurlyeq \sigma^2/nQ_{J^cJ^c}.$$

This implies that with probability larger than $1 - 3P(\Omega^*(r) > \lambda\rho\eta/2)$, we have

$$\Omega_J^*(r_J) \leqslant \lambda/2 \qquad \text{and} \qquad (\Omega^J)^*(r_{J^c} - Q_{J^cJ}Q_{JJ}^{-1}r_J) \leqslant \lambda\eta/2.$$

We denote by $\tilde{w}$ the unique (because Q_{JJ} is invertible) minimum of $\frac{1}{2n}\|y - Xw\|_2^2 + \lambda\Omega(w)$, subject to $w_{J^c} = 0$. $\tilde{w}_J$ is defined through $Q_{JJ}(\tilde{w}_J - {w_J}^*) - r_J = -\lambda s_J$ where $s_J \in \partial\Omega_J(\tilde{w}_J)$ (which implies that $\Omega_J^*(s_J) \leqslant 1$) , i.e., $\tilde{w}_J - w_J^* = Q_{JJ}^{-1}(r_J - \lambda s_J)$. We have:

$$\begin{aligned}
\|\tilde{w}_J - w_J^*\|_\infty &\leqslant \max_{j\in J} |\delta_j^\top Q_{JJ}^{-1}(r_J - \lambda s_J)| \\
&\leqslant \max_{j\in J} \Omega_J(Q_{JJ}^{-1}\delta_j)\Omega_J^*(r_J - \lambda s_J)| \\
&\leqslant \max_{j\in J} \|Q_{JJ}^{-1}\delta_j\|_p F(J)^{1-1/p}[\Omega_J^*(r_J) + \lambda\Omega_J^*(s_J)] \\
&\leqslant \max_{j\in J} \kappa^{-1}|J|^{1/p}F(J)^{1-1/p}[\Omega_J^*(r_J) + \lambda\Omega_J^*(s_J)] \leqslant \frac{3}{2}\lambda|J|^{1/p}F(J)^{1-1/p}\kappa^{-1}.
\end{aligned}$$

Thus if $2\lambda|J|^{1/p}F(J)^{1-1/p}\kappa^{-1} \leqslant \nu$, then $\|\tilde{w} - w^*\|_\infty \leqslant \frac{3\nu}{4}$, which implies $\mathrm{Supp}(\tilde{w}) \supset \mathrm{Supp}(w^*)$.

In the neighborhood of $\tilde{w}$, we have an exact decomposition of the norm, hence, to show that $\tilde{w}$ is the unique global minimum, we simply need to show that since we have $(\Omega^J)^*(r_{J^c} - Q_{J^cJ}Q_{JJ}^{-1}r_J) \leqslant \lambda\eta/2$, $\tilde{w}$ is the unique minimizer of Eq. (10). For that it suffices to show that $(\Omega^J)^*(Q_{J^cJ}(\tilde{w}_J - w_J^*) - r_{J^c}) < \lambda$. We have:

$$\begin{aligned}
(\Omega^J)^*(Q_{J^cJ}(\tilde{w}_J - w_J^*) - r_{J^c}) &= (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}(r_J - \lambda s_J) - r_{J^c}) \\
&\leqslant (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}r_J - r_{J^c}) + \lambda(\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}s_J) \\
&\leqslant (\Omega^J)^*(Q_{J^cJ}Q_{JJ}^{-1}r_J - r_{J^c}) + \lambda(\Omega^J)^*[(\Omega_J(Q_{JJ}^{-1}Q_{Jj}))_{j\in J^c}] \\
&\leqslant \lambda\eta/2 + \lambda(1-\eta) < \lambda,
\end{aligned}$$

which leads to the desired result. □

F.2 Proof of proposition 6 (Consistency)

Proof. Like for the proof of Proposition 5, we have

$$\Omega(x) \geqslant \Omega_J(x_J) + \Omega^J(x_{J^c}) \geqslant \Omega_J(x_J) + \rho\,\Omega_{J^c}(x_{J^c}) \geqslant \rho\,\Omega(x).$$

Thus, if we assume $\Omega^*(q) \leqslant \lambda\rho/2$, then $\Omega_J^*(q_J) \leqslant \lambda/2$ and $(\Omega^J)^*(q_{J^c}) \leqslant \lambda/2$. Let $\Delta = \hat{w} - w^*$.

We follow the proof from Bickel et al. (2009) by using the decomposition property of the norm Ω. We have, by optimality of $\hat{w}$:

$$\frac{1}{2}\Delta^\top Q\Delta + \lambda\Omega(w^* + \Delta) + q^\top\Delta \leqslant \lambda\Omega(w^* + \Delta) + q^\top\Delta \leqslant \lambda\Omega(w^*)$$

Using the decomposition property,

$$\lambda\Omega_J((w^* + \Delta)_J) + \lambda\Omega^J((w^* + \Delta)_{J^c}) + q_J^\top\Delta_J + q_{J^c}^\top\Delta_{J^c} \leqslant \lambda\Omega_J(w_J^*),$$

$$\lambda\Omega^J(\Delta_{J^c}) \leqslant \lambda\Omega_J(w_J^*) - \lambda\Omega_J(w_J^* + \Delta_J) + \Omega_J^*(q_J)\Omega_J(\Delta_J) + (\Omega^J)^*(q_{J^c})\Omega^J(\Delta_{J^c}), \quad \text{and}$$

$$(\lambda - (\Omega^J)^*(q_{J^c}))\Omega^J(\Delta_{J^c}) \leqslant (\lambda + \Omega_J^*(q_J))\Omega_J(\Delta_J).$$

Thus $\Omega^J(\Delta_{J^c}) \leqslant 3\Omega_J(\Delta_J)$, which implies $\Delta^\top Q\Delta \geqslant \kappa\|\Delta_J\|_2^2$ (by our assumption which generalizes the usual ℓ_1-restricted eigenvalue condition). Moreover, we have:

$$\begin{aligned}
\Delta^\top Q\Delta &= \Delta^\top(Q\Delta) \leqslant \Omega(\Delta)\Omega^*(Q\Delta) \\
&\leqslant \Omega(\Delta)(\Omega^*(q) + \lambda) \leqslant \frac{3\lambda}{2}\Omega(\Delta) \text{ by optimality of } \hat{w} \\
\Omega(\Delta) &\leqslant \Omega_J(\Delta_J) + \rho^{-1}\Omega^J(\Delta_{J^c}) \\
&\leqslant \Omega_J(\Delta_J)(3 + \frac{1}{\rho}) \leqslant \frac{4}{\rho}\Omega_J(\Delta_J).
\end{aligned}$$

This implies that $\kappa\Omega_J(\Delta_J)^2 \leqslant \Delta^\top Q\Delta \leqslant \frac{6\lambda}{\rho}\Omega_J(\Delta_J)$, and thus $\Omega_J(\Delta_J) \leqslant \frac{6\lambda}{\kappa\rho}$, which leads to the desired result, given the previous inequalities.

□

F.3 Proof of proposition 7

Proof. We have $\Omega^*(z) = \max_{A\in\mathcal{D}_F} \frac{\|z_A\|_q}{F(A)^{1/q}}$. Thus, from the union bound, we get

$$\mathbb{P}(\Omega^*(z) > t) \quad \leqslant \quad \sum_{A\in\mathcal{D}_F} \mathbb{P}(\|z_A\|_q^q > t^q F(A)).$$

We can then derive concentration inequalities. We have $\mathbb{E}\|z_A\|_q \leqslant (\mathbb{E}\|z_A\|_q^q)^{1/q} = (|A|\mathbb{E}|\varepsilon|^q)^{1/q} \leqslant 2|A|^{1/q}q^{1/2}$, where ε is a standard normal random variable. Moreover, $\|z_A\|_q \leqslant \|z_A\|_2$ for $q \geqslant 2$, and $\|z_A\|_q \leqslant |A|^{1/q-1/2}\|z_A\|_2$ for $q \leqslant 2$. We can thus use the concentration of Lipschitz-continuous functions of Gaussian variables, to get for $p \geqslant 2$ and $u \geqslant 0$,

$$\mathbb{P}\big(\|z_A\|_q \geqslant 2|A|^{1/q}\sqrt{q} + u\big) \leqslant e^{-u^2/2}.$$

For $p < 2$ (i.e., $q > 2$), we obtain

$$\mathbb{P}\big(\|z_A\|_q \geqslant 2|A|^{1/q}\sqrt{q} + u\big) \leqslant e^{-u^2|A|^{1-2/q}/2}.$$

We can also bound the expected norm $\mathbb{E}[\Omega^*(z)]$, as

$$\mathbb{E}[\Omega^*(z)] \leqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}}.$$

Together with $\Omega^*(z) \leqslant \|z\|_2 \max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}$, we get

$$\mathbb{P}\left(\Omega^*(z) \geqslant 4\sqrt{q\log(2|\mathcal{D}_F|)}\max_{A\in\mathcal{D}_F}\frac{|A|^{1/q}}{F(A)^{1/q}} + u\max_{A\in\mathcal{D}_F}\frac{|A|^{(1/q-1/2)_+}}{F(A)^{1/q}}\right) \leqslant e^{-u^2/2}.$$

□